

# Objectives and Scope

---

## Course Goals

- Introduction to basic statistical terms
- Provide a structured overview of statistical concepts
- Explain interpretation and basis of statistical distributions and hypothesis testing
- Provide helpful “tricks of the trade”

## Beyond the Scope of this Training

- Comprehensive coaching on Statistics
- Derivation of statistical formulas or terms (unless required as part of methodology explanation)

## Self Study Goals

- In-depth research on advanced Statistical concepts
- Innovations and new techniques related to methodology
- Discussion on advanced concepts can be taken up offline

# Table of Contents

---

## 1. Dataset Basics

### 1.1. Data Set Dimensions

- 1.1.1. Data Set
- 1.1.2. Observation
- 1.1.3. Variable

### 1.2. Variable Type

- 1.2.1. Categorical Variable
- 1.2.2. Quantitative Variable

### 1.3. Scales of Measurement

- 1.3.1. Properties of Measurement Scales
- 1.3.2. Nominal Scale
- 1.3.3. Ordinal Scale
- 1.3.4. Interval Scale
- 1.3.5. Ratio Scale
- 1.3.6. Scales of Measurement: Summary

## 2. Univariate Analysis

### 2.1. Univariate Analysis

- 2.1.1. Plotting and Visualizing a Distribution
- 2.1.2. Examining a Distribution
- 2.1.3. Percentiles

---

## **2.2. Centre (Measures of Central Tendency)**

2.2.1. Mean

2.2.2. Median

2.2.3. Mode

## **2.3. Spread (Measures of Dispersion)**

2.3.1. Range

2.3.2. Inter-Quartile Range

2.3.3. Variance

2.3.4. Standard Deviation

2.3.5. Coefficient of Variation

## **2.4. Shape (Skewness / Kurtosis)**

2.4.1. Skewness

2.4.2. Kurtosis

## **2.5. Box Plot**

# **3. Sampling Distributions**

## **3.1. Sampling**

3.1.1. Sample Selection

3.1.2. Parameter and Statistic

3.1.3. Sampling Distribution of a Statistic

3.1.4. Standard Error

---

## **3.2. Random Variable & Probability Distributions**

- 3.2.1. Random Variable
- 3.2.2. Probability Mass Function (p.m.f.)
- 3.2.3. Probability Density Function (p.d.f.)
- 3.2.4. Cumulative Distribution Function (c.d.f.)

## **3.3. List of Distributions**

- 3.3.1. Examples of Discrete Distributions
- 3.3.2. Example of Continuous Distributions

## **4. Hypothesis Testing**

### **4.1. Key Concepts**

- 4.1.1. Statistical Hypothesis
- 4.1.2. Tests of Significance
- 4.1.3. Test Statistic
- 4.1.4. Types of Error, Level of Significance and Power of Test
- 4.1.5. Confidence Interval
- 4.1.6. Critical Regions
- 4.1.7. The p-value or Exact Level of Significance

### **4.2. Step-by-Step Process**

---

## **5. Scatter Plots and Correlations**

### **5.1. Scatter Plot**

5.1.1. Explanatory and Response Variables

5.1.2. Plotting and Visualizing a Scatter Plot

5.1.3. Examining a Scatter Plot

### **5.2. Correlation**

5.2.1. Pearson's Correlation Coefficient

5.2.2. Spearman's Rank Correlation Coefficient

## **Appendix**

**A.1.** List of Some Commonly Used Tests

**A.2.** Examples of Discrete Distributions

**A.3.** Examples of Continuous Distributions

---

## Chapter 1: Dataset Basics

---

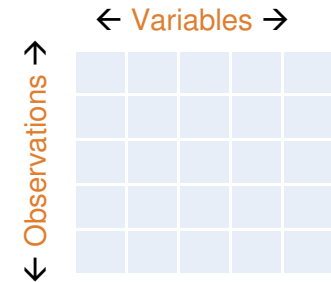
# 1.1 Data Set Dimensions

---

## 1.1.1. Data Set

A data set contains information in a set of rows and columns

- Row → Observation
- Column → Variable



## 1.1.2. Observation

Observations are the objects described by a set of data. They may be anything.

Few examples:

- Individuals (like customers of a bank, employees of a company, students of a class)
- Transactions
- Accounts
- Zip Codes

## 1.1.3. Variable

A variable is any characteristic of an observation and can take different values for different observations.

Few examples:

- A customer's tenure or an employee's salary or a student's marks
- A transaction's date, time or amount
- An account's balance or spending limit
- A zip code's population

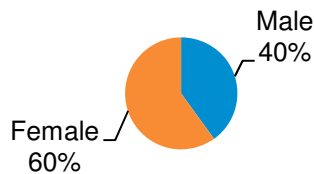
# 1.2 Variable Type

## 1.2.1. Categorical Variable

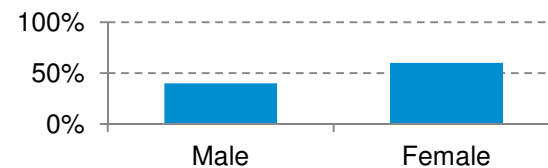
A categorical variable places each observation into one of the several groups or categories, like the gender variable with male or female categories

Such variables are best represented by a **Pie Chart** or a **Bar Graph**

**Gender Distribution**



**Gender Distribution**

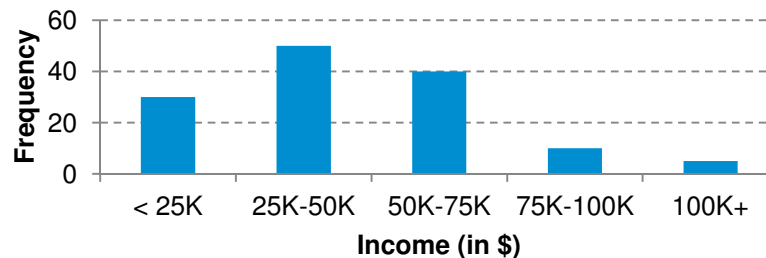


## 1.2.2. Quantitative Variable

A quantitative variable has numerical values that measure some characteristic of each observation, like height in centimeters or salary in dollars per year

Such variables are best represented by a **Histogram**

**Income Distribution**



### Things to Remember

Before drawing the histogram, the quantitative variable is divided into equal classes



# Exercise

---

**Exercise 1.** In the 'Bar Graph' of Gender Distribution (in [Section 1.2.1](#)), there are two categories. In the 'Histogram' of Income Distribution (in [Section 1.2.2](#)), there are five categories. Otherwise, how are they different from each other?

[Hint: Does the sequence of categories matter?]

# 1.3 Scales of Measurement

## 1.3.1. Properties of Measurement Scales

### Four Key Properties Relating to Scales of Measurement

Each number has a particular meaning

Numbers have an inherent order from smaller to larger

Differences between numbers (units) anywhere on the scale are the same

Zero point represents the absence of the characteristic being measured



**Identity**

**Magnitude**

**Equal Intervals**

**True Zero**

**Scales**

**Nominal Scale**

?

?

?

?

**Ordinal Scale**

?

?

?

?

**Interval Scale**

?

?

?

?

**Ratio Scale**

?

?

?

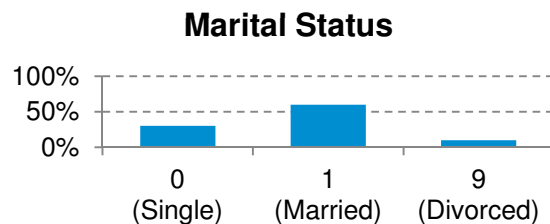
?

### 1.3.2. Nominal Scale

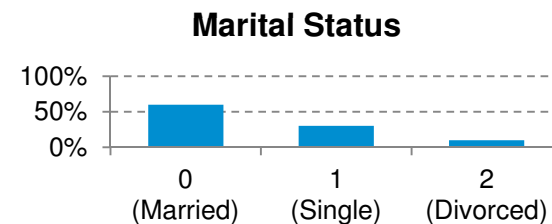
#### Example: 'Marital Status'

- Consider a variable 'Marital Status' with three categories: Single, Married and Divorced
- Arbitrarily, a number is assigned to the three categories

**Scenario A:** Single denoted by '0', Married denoted by '1' and Divorced denoted by '9'



**Scenario B:** Single denoted by '1', Married denoted by '0' and Divorced denoted by '2'



#### As Observed

- Scale points '0', '1' and '9' have specific meanings with respect to a scenario
- Scale points are mere names of the categories. The categories on nominal scale can not be rank-ordered
- Difference between two consecutive scale points is not 'equal'. In fact, calculation of such difference is not even meaningful.
- The scale point '0' is not the 'true' zero point (Here '0' does not mean 'absence' of marital status)

### 1.3.3. Ordinal Scale

#### Example: 'Preference Rank' for Car's Color

- Suppose a car buyer is asked to provide preference ranking for colors: white, black, red and blue

Color	Buyer's State of Mind	Scenario A	Scenario B
		Preference Rank (on a scale of 1 to 4)	Preference Rank (on a scale of 0 to 3)
Black	I simply love black	1	0
White	I strongly like white too	2	1
Red	Red looks very sporty	3	2
Blue	I hate blue	4	3

#### As Observed

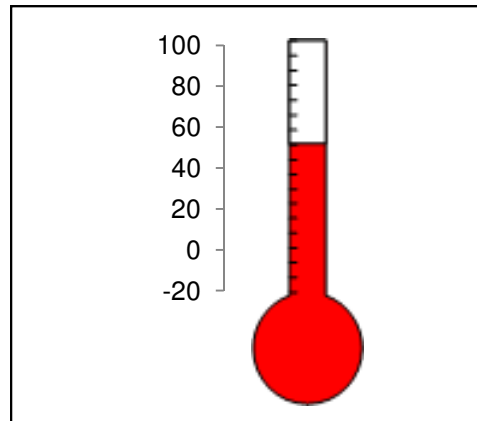
1. Every scale point (preference rank) has a particular meaning with respect to a scenario
2. Each number on the scale is different from the other and all of them can be rank-ordered
3. Rank difference between the last two preferences can not be assumed to be equivalent to rank difference between the first two preferences. Clearly, the customer's preference for black over white is not as strong as his preference for red over blue.
4. The scale point '0' is not the 'true' zero point (In Scenario B, black color has '0' rank, which does not mean the absence of preference. Rather it is the most preferred color.)

---

### 1.3.4. Interval Scale

#### Example: Temperature Measurement on Celsius Scale

- The thermometer identifies how many units of mercury correspond to the temperature measured



#### As Observed

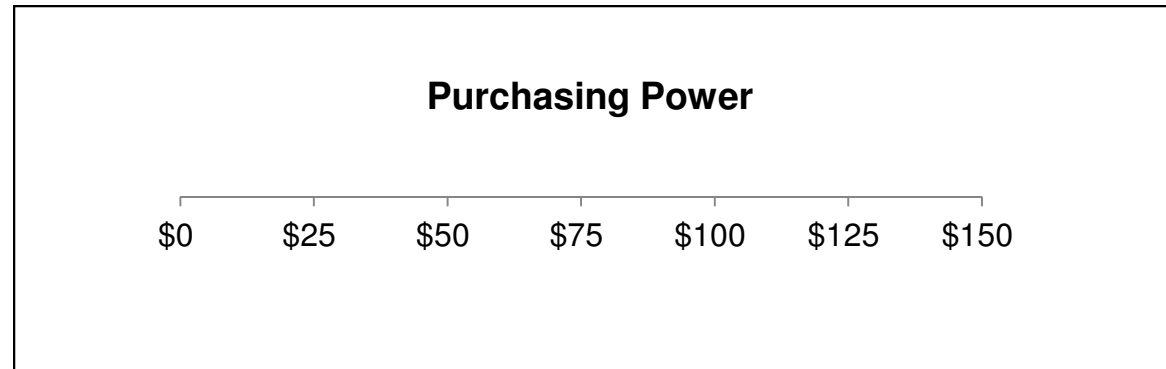
1. Every scale point has a particular meaning
2. Each number on scale is different from the other and all of them can be rank-ordered (80°C is hotter than 60°C)
3. There is the same 10-degree difference in temperature between 20° and 30° as between 50° and 60°
4. 0°C is not a 'true' zero point. It does not indicate absence of temperature; it is an arbitrary point on the scale

---

### 1.3.5. Ratio Scale

#### Example: Purchasing Power

- The purchasing power is measured in terms of money



#### As Observed

1. Every scale point has a particular meaning
2. Numbers on scale can be rank-ordered (\$150 has more purchasing power than \$125)
3. There is the same \$25 difference in purchasing power between \$100 and \$125 as between \$50 and \$75
4. \$0 is a 'true' zero point. \$0 means no money and absolutely no ability to purchase anything

### 1.3.6. Scales of Measurement: Summary

#### Four Key Properties Relating to Scales of Measurement

Each number has a particular meaning

Numbers have an inherent order from smaller to larger

Differences between numbers (units) anywhere on the scale are the same

Zero point represents the absence of the characteristic being measured



**Identity**

**Magnitude**

**Equal Intervals**

**True Zero**

**Nominal Scale**



**Ordinal Scale**



**Interval Scale**



**Ratio Scale**



**Scales**

# Exercise

---

**Exercise 2.** List down at least 3 examples each for

- a. Nominal Scale Variable
- b. Ordinal Scale Variable
- c. Interval Scale Variable
- d. Ratio Scale Variable



---

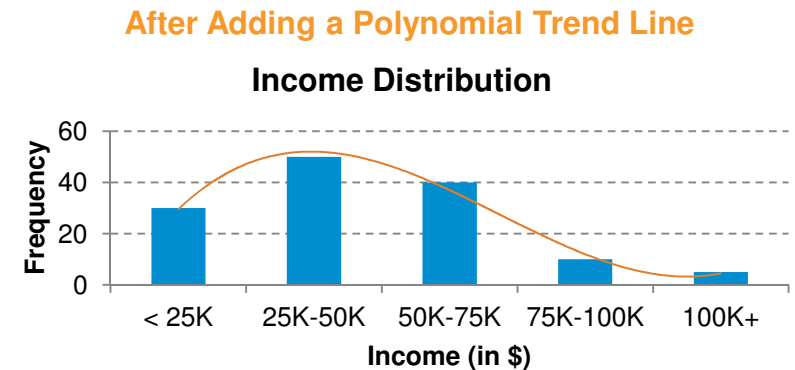
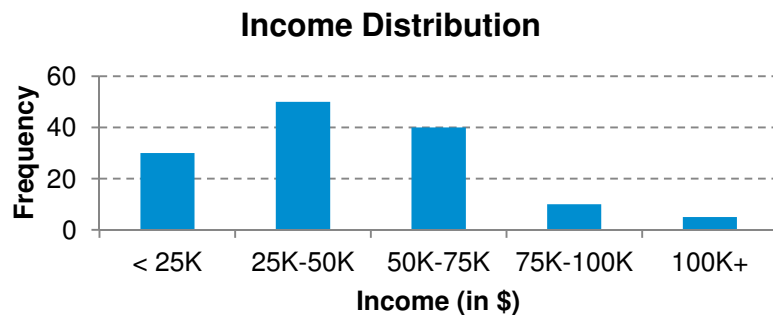
## Chapter 2: Univariate Analysis

---

## 2.1 Univariate Analysis

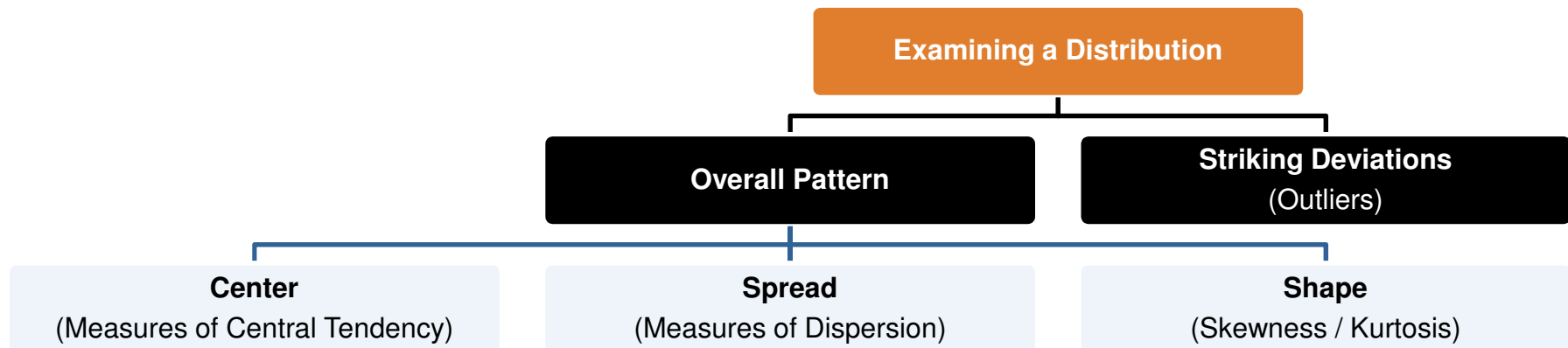
### 2.1.1. Plotting and Visualizing a Distribution

Recall the Histogram of Dollar Income plotted in [Section 1.2.2](#).



### 2.1.2. Examining a Distribution

Look for the **overall pattern** and for **striking deviations** from that pattern



### 2.1.3. Percentiles

A percentile is the value of a variable below which a certain percent of observations fall.



**SAS Tip**

**PROC UNIVARIATE**

The  $p^{\text{th}}$  percentile is a value such that

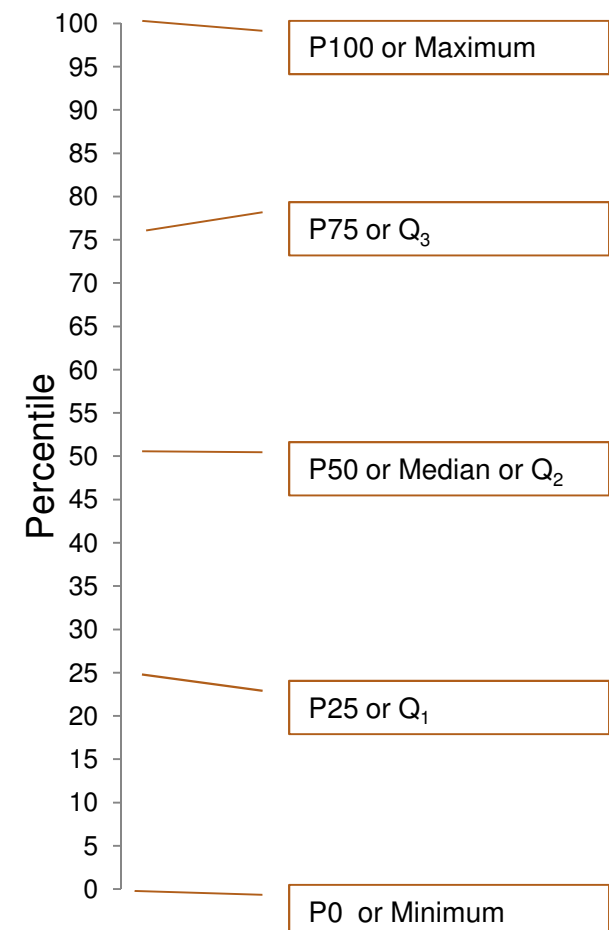
- At most (p)% of the observations are less than this value; and
- At most (1 – p)% are greater

when the data is sorted

#### Examples:

- P1 (1<sup>st</sup> percentile) cuts off the lowest 1% of sorted data
- P98 (98<sup>th</sup> percentile cuts off the lowest 98% of sorted data

Term Used	Number of Splits
Quartiles	4
Deciles	10
Pentiles	20
Percentile	100



## 2.2 Center (Measures of Central Tendency)

### 2.2.1. Mean

Mean is arithmetic average of the observations.



SAS Tip

PROC MEANS

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$n$  : Number of observations

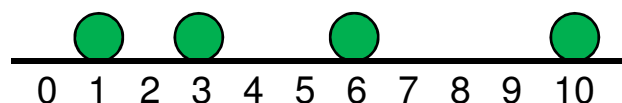
$X_i$  :  $i^{th}$  observation

#### Think about it..

If frequency of any observation is more than 1 (e.g. if two students scored the same marks), how would the formula for mean calculation look like?

- The Most Common Measure of Central Tendency
- Affected by Extreme Values (Outliers)
- 'Center of Gravity' of a Distribution - the point on which the distribution would balance

**Example:** Find average marks scored by a group of 4 students: A (1 mark), B (3 marks), C (6 marks), D (10 marks)



$$\bar{X} = \frac{1+3+6+10}{4} = \frac{20}{4} = 5$$

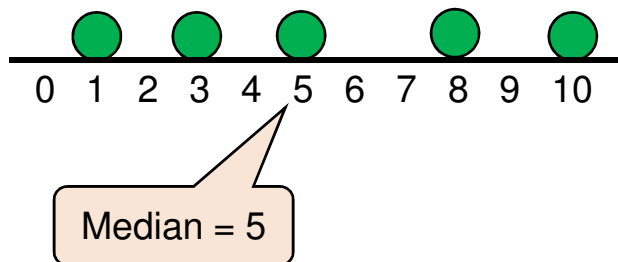
## 2.2.2. Median

Median is the midpoint of a distribution.

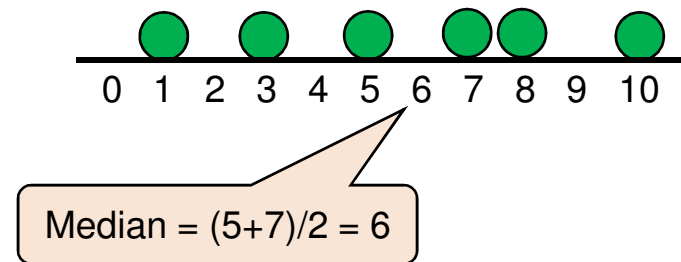
$$M = \left( \frac{n+1}{2} \right)^{th} \text{ Observation}$$

- A Robust Measure of Central Tendency
- Not Affected by Extreme Values (Outliers)

If number of observations (n) is odd, the median (M) is the center observation in the ordered list



If number of observations (n) is even, the median (M) is the mean of the two center observations in the ordered list



**SAS Tip**

**PROC MEANS**



### Things to Remember

Before calculating median, always arrange all observations in order of size, from smallest to largest. However, if you are using PROC MEANS in SAS, there is no need to sort manually. SAS does it automatically.

### 2.2.3. Mode

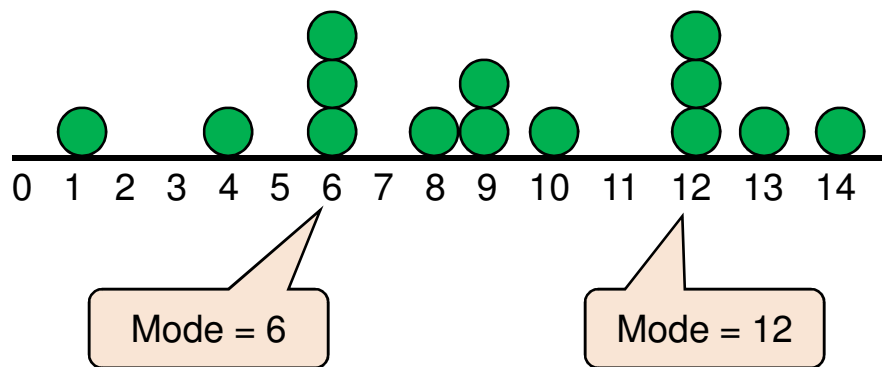
Mode is the most frequently occurring observation.



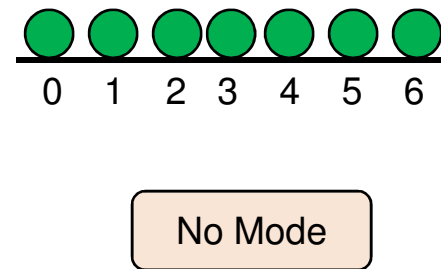
SAS Tip

PROC UNIVARIATE

There may be multiple modes



There may not be a mode



- Not Affected by Extreme Values (Outliers)
- Can be used for Numerical as well as Categorical Data

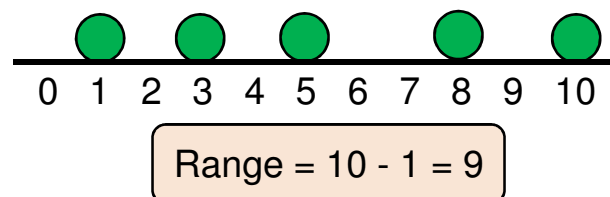
## 2.3 Spread (Measures of Dispersion)

### 2.3.1. Range

Range is the difference between the maximum and the minimum values

$$\text{Range} = \text{MaxValue} - \text{MinValue}$$

- By definition, Range is affected by Extreme Values
- Independent of data distribution



SAS Tip

PROC UNIVARIATE

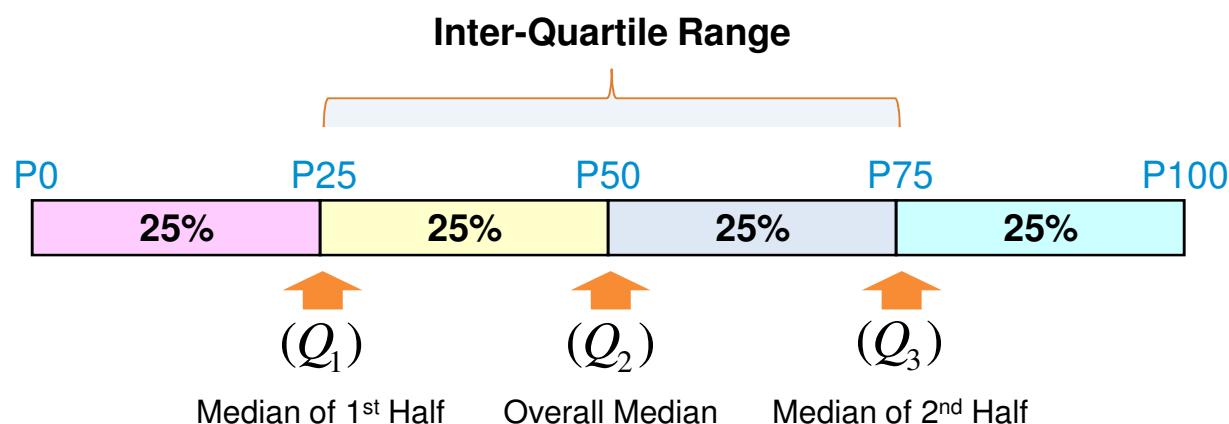
### 2.3.2. Inter-Quartile Range

Inter-Quartile Range is the difference between the first and the third quartile

- Mid-spread (Spread in the middle 50%)
- Independent of Extreme Values

$$IQR = Q_3 - Q_1$$

$$Q_i = \left( \frac{i \times (n+1)}{4} \right)^{th} \text{ Observation } n$$



---

### 2.3.3. Variance

Variance is an average of the squares of the deviations of a set of observations from their mean.

$$\sigma^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

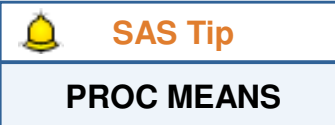
- Affected by Extreme Values (outliers)
- Variance = 0 if there is no spread (when all observations have the same value); Otherwise Variance > 0

### 2.3.4. Standard Deviation

Standard deviation is the square root of the variance.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

- Affected by Extreme Values (outliers)
- Std. Deviation = 0 if there is no spread (when all observations have the same value); Otherwise Std. Dev. > 0
- Std. Deviation has the same units of measurement as the original observations





---

### 2.3.5. Coefficient of Variation

Coefficient of variation (CV) is a normalized measure of dispersion of a probability distribution.

$$CV = \frac{\sigma}{X} \times 100 \%$$

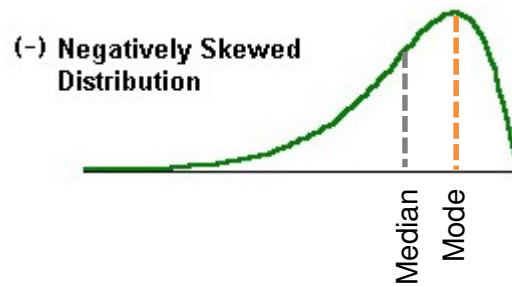
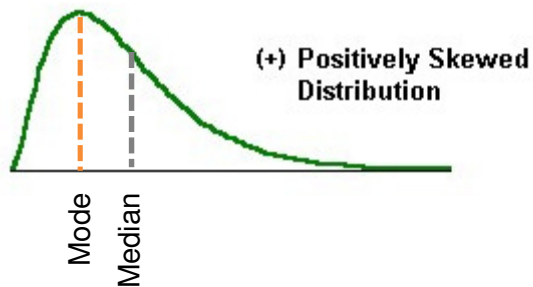
- CV is unit-less and hence can be used for comparing variation in data sets with different units
- When the mean value is close to zero, CV will approach infinity and is hence sensitive to small changes in the mean
- CV is sensitive to outliers

## 2.4 Shape (Skewness / Kurtosis)

### 2.4.1. Skewness

Skewness is the degree of departure from symmetry of a distribution.

■ Positively Skewed Distribution	: Tail pulled in the positive direction
■ Negatively Skewed Distribution	: Tail pulled in the negative direction



SAS Tip

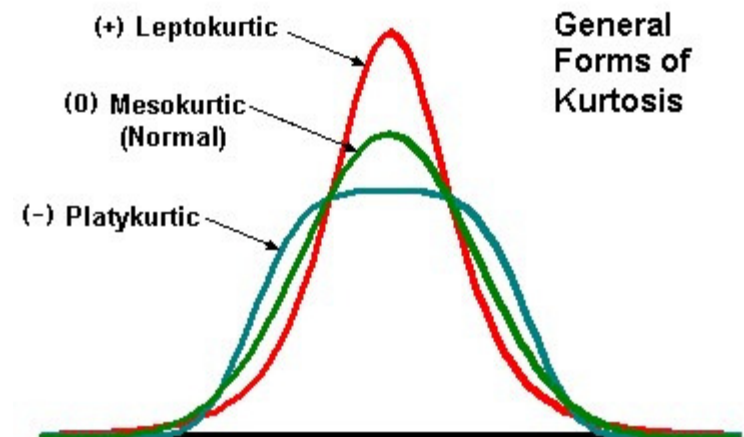
PROC UNIVARIATE

### 2.4.2. Kurtosis

Kurtosis is the degree of peakedness of a distribution.

■ Mesokurtic Distribution	: Normal distribution
■ Leptokurtic Distribution	: Higher peak and heavier tails
■ Platykurtic Distribution	: Lower peak and lighter tails

Mesokurtic Curve	: Kurtosis = 3
Leptokurtic Curve	: Kurtosis > 3
Platykurtic Curve	: Kurtosis < 3



# Example

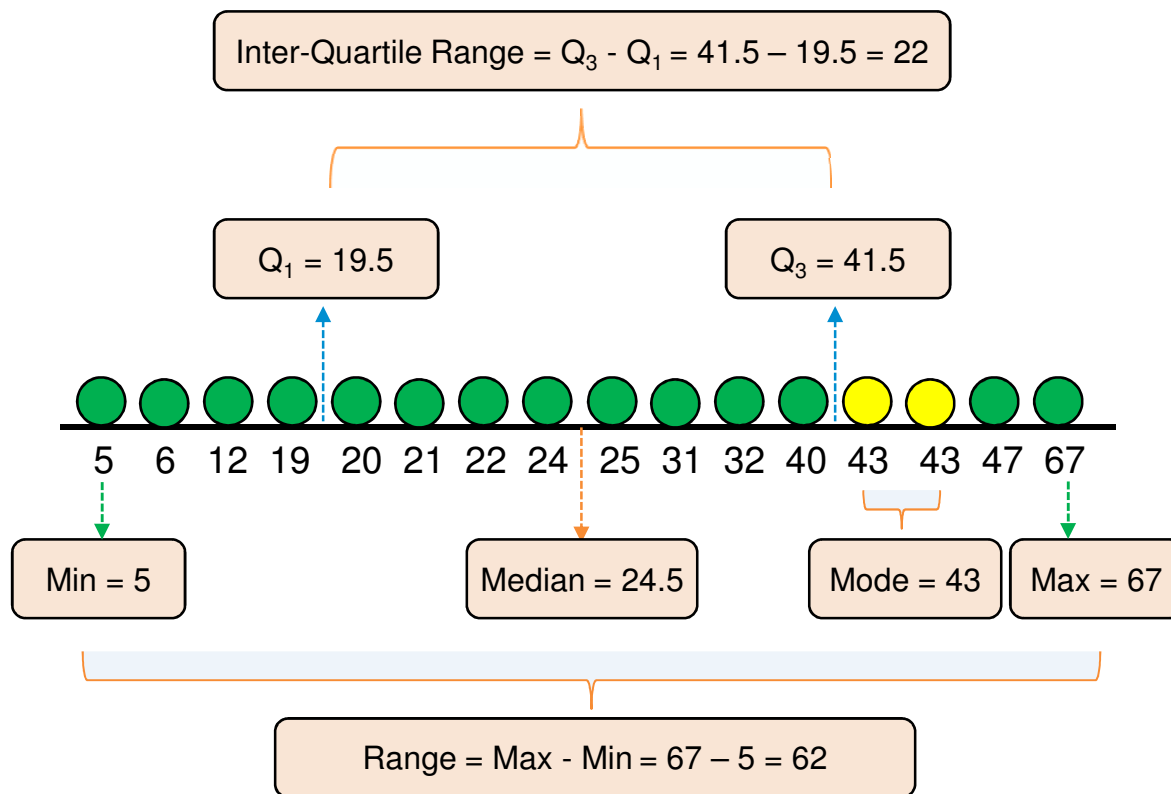
**Example:** For 16 students, marks in Mathematics are given below:

40, 31, 19, 22, 47, 25, 24, 32, 43, 12, 20, 5, 6, 21, 43, 67

Let us calculate measures of central tendency and dispersion, without using SAS.

Arrange marks in ascending order:

5, 6, 12, 19, 20, 21, 22, 24, 25, 31, 32, 40, 43, 43, 47, 67



	A	B	C
1	Marks (X)	X - Mean	(X-Mean) <sup>2</sup>
2	5	-23.6	555.2
3	6	-22.6	509.1
4	12	-16.6	274.3
5	19	-9.6	91.4
6	20	-8.6	73.3
7	21	-7.6	57.2
8	22	-6.6	43.1
9	24	-4.6	20.8
10	25	-3.6	12.7
11	31	2.4	5.9
12	32	3.4	11.8
13	40	11.4	130.8
14	43	14.4	208.4
15	43	14.4	208.4
16	47	18.4	339.9
17	67	38.4	1477.4
18	$\Sigma X = 457$	$\Sigma(X - \text{Mean}) = 0.0$	$\Sigma(X - \text{Mean})^2 = 4019.9$

$$\bar{X} = \frac{457}{16} = 28.56 \quad \sigma = \sqrt{\frac{4019.9}{(16-1)}} = 16.37 \quad CV = \frac{16.37}{28.56} = 57\%$$

# Example

For the same example, let us compute measures of center, spread and shape using SAS.

## Convert Raw Data to SAS Format

```
data outlib.marks_data;  
infile datalines;  
input marks;  
datalines;  
40  
31  
19  
22  
47  
25  
24  
32  
43  
12  
20  
5  
6  
21  
43  
67  
;
```

## Run PROC UNIVARIATE on SAS Data Set

```
proc univariate data = outlib.marks_data modes;  
var marks;  
output out = outlib.marks_univariate  
n = n  
min = min  
mean = mean  
median = median  
mode = mode  
max = max  
range = range  
q1 = q1  
q3 = q3  
qrange = qrange  
std = std_dev  
skewness = skewness  
kurtosis = kurtosis  
;  
run;
```

Output Data Set

## Output

marks_univariate.sas7bdat													
	n	mean	std_dev	skewness	kurtosis	max	q3	median	q1	min	range	qrange	mode
1	16	28.5625	16.37058	0.68004	0.542626	67	41.5	24.5	19.5	5	62	22	43

## 2.5 Box Plot

A **box-plot** is a graph of the summary of a variable's distribution

- A central box spans the quartiles  $Q_1$  and  $Q_3$
- A line in the box marks the median  $M$
- A symbol (+) marks the mean
- Lines extend from the box out to the min. and max. values

Let us create the box plot graph for the previous example.

```
univariate.sas

data outlib.marks_data;
set outlib.marks_data;
format Boxplot $1.;
Boxplot = "";
run;

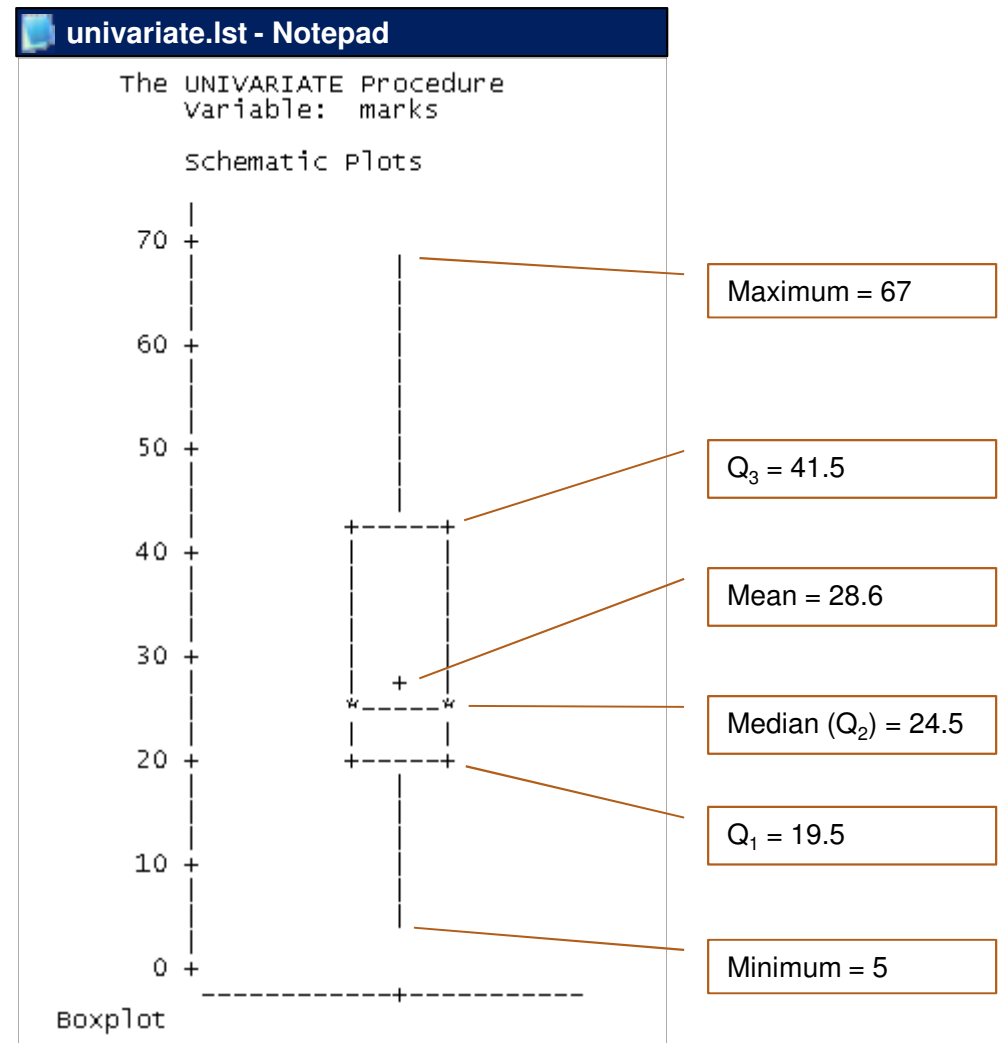
ods select ssplots ;
proc univariate data = outlib.marks_data;
var marks;
by Boxplot;
run;
```



### Some More Information

Above snippet works well in Base SAS.

For using PROC BOXPLOT, you will need SAS/GRAPH



# Exercise

---

**Exercise 3.** For 15 randomly chosen employees working in Delhi NCR, per month salary (in thousand rupees) is given below:

4, 25, 30, 30, 30, 31, 32, 35, 50, 50, 50, 55, 60, 74, 110

- a. In MS Excel spreadsheet (without using inbuilt functions), calculate
  - i. Mean
  - ii. Median
  - iii. Mode
  - iv. Range
  - v. Inter-Quartile Range
  - vi. Variance
  - vii. Standard Deviation
  - viii. Coefficient of Variation
- b. In SAS (using SAS procedures), compute and report all of the above metrics along with Skewness and Kurtosis
- c. Using SAS, create box-plot graph

**Exercise 4.** In a class of 50 students, all scored 70 marks in Statistics. Calculate Mean, Median, Mode, Range, Inter-Quartile Range, Variance, Standard Deviation and Coefficient of Variation.

[Hint: Do you really need a calculator or any software to answer this?]

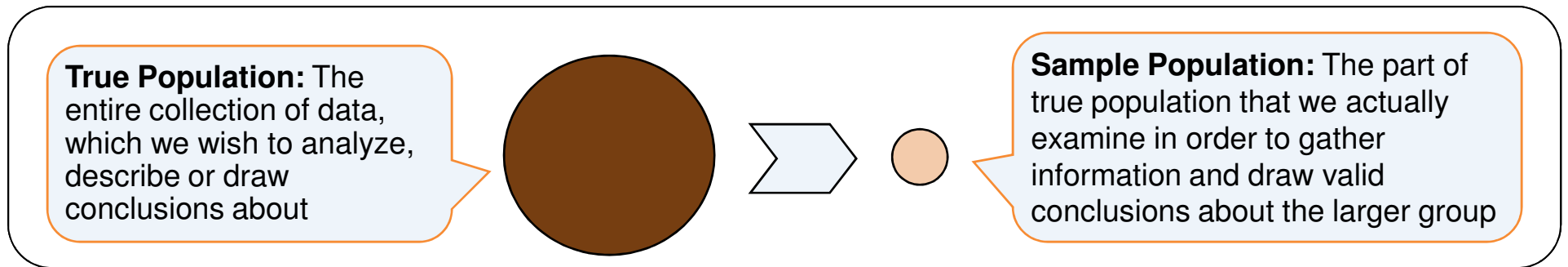
---

## Chapter 3: Sampling Distributions

---

# 3.1 Sampling

## 3.1.1. Sample Selection



### Two Most Frequently Used Methods of Sample Selection



**SAS Tip**

**PROC SURVEYSELECT**

#### **Simple Random Sampling:**

- This method selects units with equal probability and without replacement
- Each possible sample of  $n$  different units out of  $N$  has the same probability of being selected
- Selection probability for each unit =  $n / N$

#### **Stratified Random Sampling:**

- This method selects random samples independently within strata
- Selection probability for a unit in stratum  $h$  =  $n_h / N_h$

**Other Methods include Sequential Random Sampling, Systematic Random Sampling etc.**



---

### 3.1.2. Parameter and Statistic

A parameter is a value, usually unknown (and which therefore has to be estimated), that has to be estimated with minimum error using statistic from one or more samples.

Parameter	Statistic
<ul style="list-style-type: none"><li>■ Describes a characteristic of a population</li><li>■ Fixed Value within the population</li><li>■ Generally represented as Greek alphabets (e.g. <math>\Sigma</math>)</li></ul>	<ul style="list-style-type: none"><li>■ Describes a characteristic of a sample</li><li>■ Value varies from sample to sample within the same population</li><li>■ Generally represented as English alphabets (e.g. S)</li></ul>

#### Example:

A properly chosen sample of 1600 people across the country was asked if they regularly watch a certain television program, and 24% said yes.

**Parameter** : The true proportion of all people in the country who watch the program

**Statistic** : 24% (as obtained from the sample of 1600 people)

### 3.1.3. Sampling Distribution of a Statistic

The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size (n) from the same population.

**Sampling distribution** is nothing but the **probability distribution of the statistic** - It tells what values a statistic takes and how often it takes those values in repeated sampling.

**Example:** A box consists of 10 counters. The color of the counter can be

- Yellow (Score Value = 1)
- Green (Score Value = 2)
- Red (Score Value = 5)
- Blue (Score Value = 4)

Suppose there are 2 yellow, 2 green, 3 red and 3 blue counters in the box. A person is asked to draw 2 counters from this box.

All possibilities of drawing 2 counters


Frequency Distribution

		1		4		6		6	$1+4+6+6 = 17$
		1		6		6			$1+6+6=13$
		3		9					$3+9=12$
		3							3

Determine the sampling distribution of the mean score of this person.

$17 + 13 + 12 + 3 = 45$

### Frequency Distribution and Mean of Scores Obtained in Each Sample

1	1 1	1	1 2	4	1 5	6	1 4	6	1+4+6+6 = 17
		1		1.5		3		2.5	
2	2 2	2	2 5	6	2 4	6			1+6+6=13
		2		3.5		3			
5	5 5	3	5 4	9					3+9=12
		5		4.5					
4	4 4	3							3
		4							

Frequency  
Mean Scores

$$17 + 13 + 12 + 3 = 45$$

If a random sample of size 'n' is drawn from a population with mean  $\mu$  and standard deviation  $\sigma$ , then the sampling distribution of the means has mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$

### Sampling Distribution of Mean Score

Mean	Frequency	Probability
1	1	1/45
1.5	4	4/45
2	1	1/45
2.5	6	6/45
3	12	12/45
3.5	6	6/45
4	3	3/45
4.5	9	9/45
5	3	3/45
<b>Sum</b>	<b>45</b>	<b>1</b>

### 3.1.4. Standard Error

Standard error is the standard deviation of the sampling distribution of the statistic. If statistic is the sample mean, and the samples are uncorrelated, the standard error is given by:

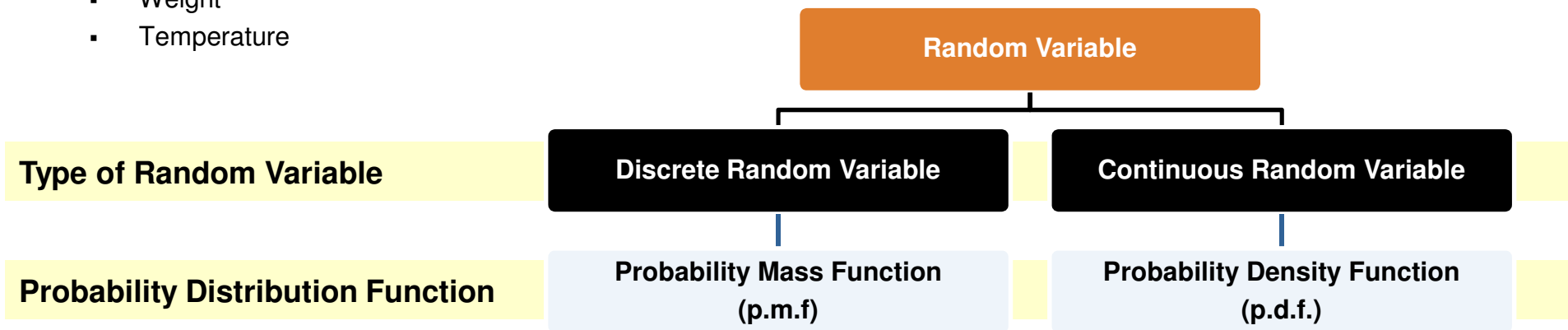
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

## 3.2 Random Variable & Probability Distributions

### 3.2.1. Random Variable

A random variable is a variable whose value is a numerical outcome of a random phenomenon. It may be discrete or continuous.

1. **Discrete Random Variables:** Random variables that have a finite (countable) list of possible outcomes, with probabilities assigned to each of these outcomes. Examples:
  - Number of Cars Owned (0, 1, 2, 3, ...)
  - Attendance (in terms of numbers of days) in a month (1, 2, 3, ... , 31)
2. **Continuous Random Variables:** Random variables that can take on any value in an interval, with probabilities given as areas under a density curve. Example:
  - Weight
  - Temperature



The **probability distribution of a random variable** is the mathematical function describing the possible values of a random variable and their associated probabilities

### 3.2.2. Probability Mass Function (p.m.f.)

A probability mass function (p.m.f.) is a function that gives the probability that a **discrete random variable**  $X$  is exactly equal to some value  $x$ .

Probability Mass Function :  $f(x) = P(X = x)$  such that the following two conditions satisfy :

1.  $0 \leq f(x) \leq 1$  for every  $x \in S$
2.  $\sum_{x \in S} f(x) = 1$

where  $S$  is the Sample Space

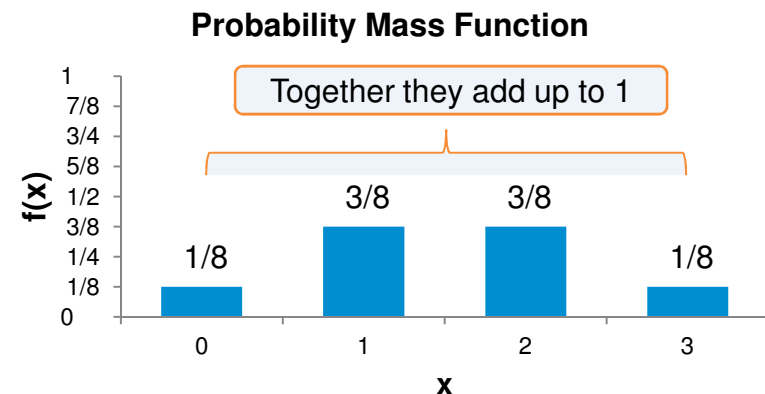
**Coin Toss Example:** Toss three unbiased coins and observe 'Number of Heads'.

Random Variable: Number of Heads Observed (which may take four values: 0, 1, 2 or 3)

Sample Space = {TTT, HTT, THT, TTH, HHT, HTH, THH, HHH}

#### Probability Mass Function

$x$	$f(x)$
0	$f(0) = P(X = 0) = P(\text{TTT}) = 1/8$
1	$f(1) = P(X = 1) = P(\text{HTT, THT, TTH}) = 3/8$
2	$f(2) = P(X = 2) = P(\text{HHT, HTH, THH}) = 3/8$
3	$f(3) = P(X = 3) = P(\text{HHH}) = 1/8$
Sum	$\Sigma f(X) = 1/8 + 3/8 + 3/8 + 1/8 = 1$



### 3.2.3. Probability Density Function (p.d.f.)

A probability density function is a function that describes the relative likelihood for a **continuous random variable**  $X$  to take values within a particular interval  $[a, b]$ .

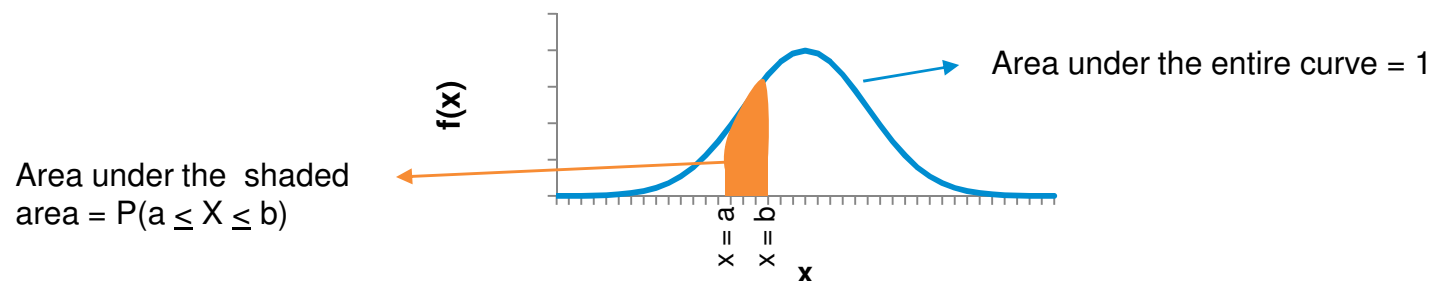
Probability Density Function :  $P(a \leq X \leq b) = \int_a^b f(x)dx$  such that the following two conditions satisfy :

1.  $f(x)$  is a non - negative integrable function
2.  $\int_{-\infty}^{\infty} f(x)dx = 1$

#### In case of a continuous random variable

- There are infinite number of outcomes and hence a probability can not be assigned to each individual
- Probabilities are assigned to intervals of outcomes by using areas under density curves
- A density curve has area exactly 1 underneath it, corresponding to total probability 1

#### Probability Density Function



### 3.2.4. Cumulative Distribution Function (c.d.f.)

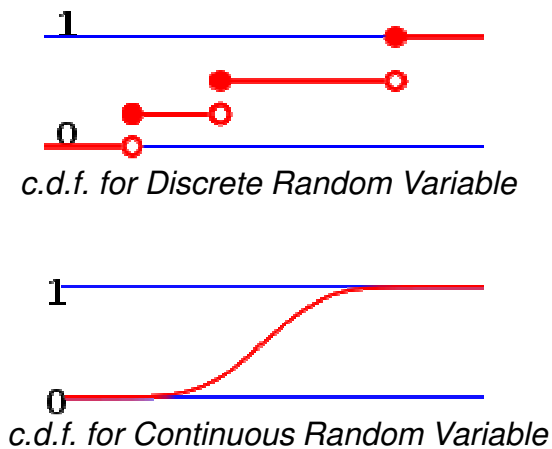
A cumulative distribution function describes the probability that a random variable  $X$  with a given probability distribution will be found at a value less than or equal to  $x$ . Intuitively, it is the "area so far" function of the probability distribution.

Cumulative Distribution Function:  $F(x) = P(X \leq x)$

For Discrete Random Variable:  $F(x_i) = P(X \leq x_i) = \sum_{k=1}^i P(X = x_k)$  [Note:  $P(X = x_k) = F(x_k) - F(x_{k-1})$ ]

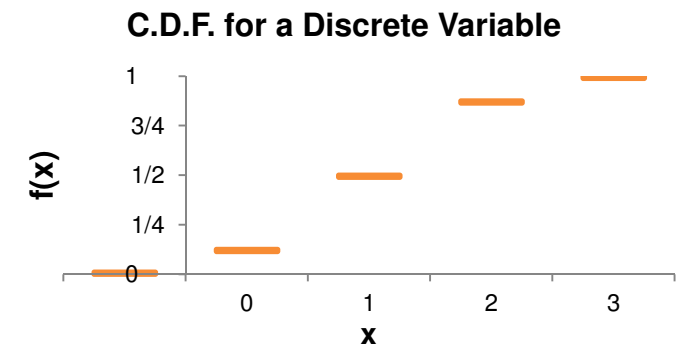
For Continuous Random Variable:  $F(x_i) = P(X \leq x_i) = \int_{-\infty}^{x_i} f(t) dt$  [Note:  $f$  is probability density function]

Properties of CDF:  $0 \leq F(x) \leq 1$ ,  $\lim_{x \rightarrow -\infty} F(x) = 0$ ,  $\lim_{x \rightarrow +\infty} F(x) = 1$



#### Coin Toss Example Continued...

x	f(x)	F(x)
0	$f(0) = 1/8$	$F(1) = 1/8$
1	$f(1) = 3/8$	$F(2) = 1/8 + 3/8 = 4/8$
2	$f(2) = 3/8$	$F(3) = 4/8 + 3/8 = 7/8$
3	$f(3) = 1/8$	$F(4) = 7/8 + 1/8 = 1$
$\Sigma f(X) = 1$		



## 3.3 List of Distributions

---

### 3.3.1. Examples of Discrete Distributions

1. Discrete Uniform Distribution [\[Details\]](#)
2. Binomial Distribution [\[Details\]](#)
3. Hypergeometric Distribution [\[Details\]](#)
4. Poisson Distribution [\[Details\]](#)
5. Geometric Distribution [\[Details\]](#)
6. Negative Binomial Distribution [\[Details\]](#)

**Note:** 'Details' link leads to Appendix A.2

### 3.3.2. Examples of Continuous Distributions

1. Continuous Uniform Distribution [\[Details\]](#)
2. Normal Distribution [\[Details\]](#)
3. Chi-Square Distribution [\[Details\]](#)
4. F Distribution [\[Details\]](#)
5. Student's t Distribution [\[Details\]](#)
6. Exponential Distribution [\[Details\]](#)
7. Gamma Distribution [\[Details\]](#)
8. Beta Distribution [\[Details\]](#)

**Note:** 'Details' link leads to Appendix A.3



# Exercise

---

**Exercise 5.** In an excel spreadsheet, plot

1. Probability Mass Function of Binomial Distribution
2. Probability Mass Function of Hypergeometric Distribution
3. Probability Mass Function of Poisson Distribution
4. Probability Mass Function of Negative Binomial Distribution
5. Probability Density Function of Normal Distribution
6. Probability Density Function of Exponential Distribution
7. Probability Density Function of Gamma Distribution

[Hint A: Explore Excel functions: BINOMDIST, HYPGEOMDIST, POISSON, NEGBINOMDIST, NORMDIST, EXPONDIST and GAMMADIST]

[Hint B: Refer Appendix A.2 and A.3]

---

## Chapter 4: Hypothesis Testing

---

# 4.1 Key Concepts

---

## 4.1.1. Statistical Hypothesis

A statistical hypothesis is some assumption or statement about a population parameter.

- This assumption may or may not be true
- It is tested on the basis of the evidence from a random sample

**Example:** Average sales amount (i.e. sales amount per transaction) of a grocery store is \$150.

Statistical hypotheses are of two types:

A. **Null Hypothesis**

Hypothesis which is being tested. It is denoted by  $H_0$ .

B. **Alternative Hypothesis**

Counter-proposition to the null hypothesis. It is denoted by  $H_1$  or  $H_A$ .

**Example:** Let the average sales amount of a grocery store be denoted by  $\mu$ . There could be three scenarios.

- |                              |                        |     |                      |
|------------------------------|------------------------|-----|----------------------|
| 1. Left-Tailed (One Tailed)  | : $H_0 : \mu \geq 150$ | and | $H_1 : \mu < 150$    |
| 2. Right-Tailed (One Tailed) | : $H_0 : \mu \leq 150$ | and | $H_1 : \mu > 150$    |
| 3. Two-Tailed                | : $H_0 : \mu = 150$    | and | $H_1 : \mu \neq 150$ |

### Can We Accept the Null Hypothesis?

Based on a sample, we can not conclude that  $H_0$  is true. So, instead of concluding “we accept  $H_0$ ”, we say “we fail to reject  $H_0$ ”.

**Note:** “Acceptance of  $H_0$ ” applies  $H_0$  is true, while “Failure to reject  $H_0$ ” implies evidence from sample is not sufficient for rejecting  $H_0$ .

## 4.1.2. Tests of Significance

Test of significance is, in general, a procedure to assess the significance of difference between two or more values.

**List of Some Commonly Used Tests:** See Appendix A.1

**Example:** The sales amount of a grocery store is normally distributed with standard deviation of \$30.2. A sample of 40 sales receipts has an average sales amount of \$137. Based on sample information, one may want to test whether the mean of sales at the grocery store is different from \$150.

In this case,

- Assumed population mean ( $\mu_0$ ) = 150 and Sample mean ( $\bar{x}$ ) = 137
- Objective is to test the significance of difference between  $\mu_0$  and  $\bar{x}$
- Population is normally distributed and its variance is known

Therefore, Z-Test should be applied

- ✓ If the difference is **not** found to be **significant**, we **do not reject  $H_0$**  and it is attributed to **pure chance due to fluctuations in sampling**
- ✓ If the difference is found to be **significant**, we **reject  $H_0$**  and it is attributed to **some non-random cause**

---

### 4.1.3. Test Statistic

A test statistic is a value calculated from a sample to test the validity of null hypothesis. Since a test statistic is a random variable, it has a probability distribution.

#### Example:

Null Hypothesis : Average sales amount of grocery store = \$150 [ i.e.  $H_0 : \mu = 150$  ]

Alternative Hypothesis : Average sales amount of grocery store  $\neq$  \$150 [ i.e.  $H_1 : \mu \neq 150$  ]

Additional Information :

1. Sales amount is normally distributed with standard deviation of \$30.2 [ i.e.  $X \sim N(\mu, 30.2^2)$  ]

2. From a sample of 40 transactions, average sales amount = \$137 [ i.e.  $\bar{x} = 137$  ]

Test Statistic : 
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{137 - 150}{30.2/\sqrt{40}} = -2.722$$

#### Properties

All good test statistics should have two properties:

- (a) They should tend to behave differently when  $H_0$  is true from when  $H_1$  is true; and
- (b) Their probability distribution should be calculable under the assumption that  $H_0$  is true. It is also desirable that tables of this probability distribution exist

#### 4.1.4. Types of Error, Level of Significance and Power of Test

There can be two types of errors in Hypothesis Testing

- **Type I Error** : Error of rejecting  $H_0$  when it is, in fact, true
- **Type II Error** : Error of not rejecting  $H_0$  when it is, in fact, false

**Size of the Test (or Level of Significance) =  $\alpha$**

- Probability of committing a Type I Error
- Maximize size of Type I Error at risk

		Situation	
		$H_0$ is true	$H_0$ is false
Decision	$H_0$ is not rejected	Right Decision	Type II Error $P(\text{Type I Error}) = \beta$
	$H_0$ is rejected	Type I Error $P(\text{Type I Error}) = \alpha$	Right Decision

**Power of the Test =  $1-\beta$**

- Probability of not committing a Type II Error
- Ability of the Test to Reject a False Null Hypothesis

- For any sample size, it is not possible to minimize both types of errors simultaneously
- In practice, a Type I error is likely to be more serious than a Type II error
- Classical Approach:
  1. Keep the probability of committing a type I error at a fairly low level (0.01, 0.05 or 0.10)
  2. Then try to minimize the probability of having a type II error (i.e. maximize the power of the test)

**Set  $\alpha$  to 1%, 5% or 10% and then try to minimize  $\beta$**

### 4.1.5. Confidence Interval

Confidence interval is the interval in which we can be 100(1 -  $\alpha$ )% sure that values of a particular statistic will lie.

Note:  $\alpha$  is the level of significance.

Significance Level	Confidence Level
0.01	99%
0.05	95%
0.10	90%

#### 95% Confidence Interval for Mean ( $\mu$ ) of a Large Sample:

$\mu$  : Population Mean

$n$  : Sample Size

$\sigma$  : Population Standard Deviation

$\bar{x}$  : Sample Mean

$$X_i \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow Z_i = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

From Normal Distribution Table,

$$P(-1.96 \leq Z_i \leq 1.96) = 0.95$$

$$\Rightarrow P(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

$$\Rightarrow P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

#### Interpretation of 95% Confidence Interval:

A 95% Confidence Interval is interpreted as the range of values within which the population parameter will fall 95% of the time, if the procedure of calculating confidence intervals is repeated on multiple samples

### 4.1.6. Critical Regions

Confidence interval is called the **acceptance region** and the areas outside it are called the **critical regions** or **regions of rejection** of the null hypothesis. The lower and upper limits of the acceptance region are called the **critical values**.

**Example:** Continuing with Grocery store example:

$H_0 : \mu = 150, H_1 : \mu \neq 150, X \sim N(\mu, 30.2^2), \bar{x} = 137, n$

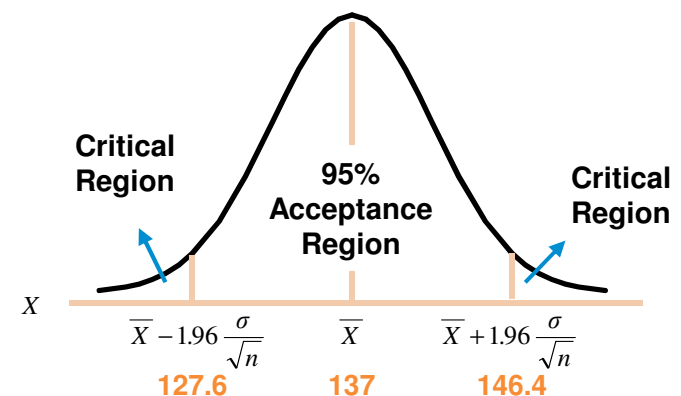
$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Rightarrow P\left(137 - 1.96 \times \frac{30.2}{\sqrt{40}} \leq \mu \leq 137 + 1.96 \times \frac{30.2}{\sqrt{40}}\right) = 0.95$$

$$\Rightarrow P(137 - 9.4 \leq \mu \leq 137 + 9.4) = 0.95$$

$$\Rightarrow P(127.6 \leq \mu \leq 146.4) = 0.95$$

95% Confidence Interval for  $\mu$  : [127.6, 146.4]



**Observation** :  $\mu = 150$  lies in the critical region

**Conclusion** :  $H_0$  is rejected at 5% significance level



### 4.1.7. The p-Value or Exact Level of Significance

Instead of preselecting  $\alpha$  (the level of significance) at arbitrary levels, such as 1%, 5% or 10%, one can obtain the p (probability) value, or exact level of significance of a test statistic. The p value is defined as the lowest significance level at which a null hypothesis can be rejected.

#### Example:

Null Hypothesis : Average sales amount of grocery store = \$150

[ i.e.  $H_0 : \mu = 150$  ]

Alternative Hypothesis : Average sales amount of grocery store  $\neq$  \$150

[ i.e.  $H_1 : \mu \neq 150$  ]

Additional Information :

1. Sales amount is normally distributed with standard deviation of \$30.2

[ i.e.  $X \sim N(\mu, 30.2^2)$  ]

2. From a sample of 40 transactions, average sales amount = \$137

[ i.e.  $\bar{x} = 137$  ]

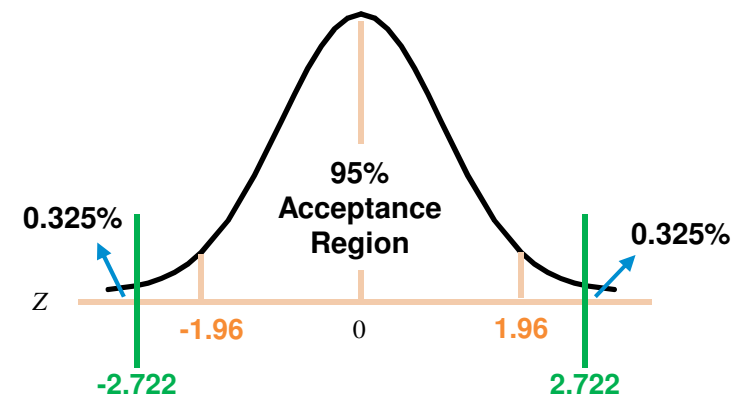
Test Statistic : 
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{137 - 150}{30.2/\sqrt{40}} = -2.722$$

It is a two-tailed test. From table, p-value = 0.0065

#### Interpretation:

p-value = 0.65% is the exact level of significance.

- Since 0.65% < 1%, we reject  $H_0$  not only at 5% level, but even at 1% level of significance
- That is, we reject  $H_0$  not with just 95% confidence, but with 99% confidence.



$P(Z \leq -1.96) = 2.5\%$ ,  $P(Z \geq 1.96) = 2.5\%$ : Adding up to 5%

$P(Z \leq -2.722) = 0.325\%$ ,  $P(Z \geq 2.722) = 0.325\%$ : Adding up to 0.65%

## 4.2 Step-by-Step Process

---

### Steps in Hypothesis Testing

Process of hypothesis testing can be summarized in following five steps:

- Step 1:** Set up Null Hypothesis ( $H_0$ ) and Alternative Hypothesis ( $H_1$ ), keeping in mind if it is going to be a single (left or right) tailed or a two-tailed test
- Step 2:** Calculate the test statistic and determine its probability distribution
- Step 3:** Choose the appropriate (1%, 5% or 10%) level of significance ( $\alpha$ )
- Step 4:** Evaluate the test statistic by either confidence interval approach or by p value approach
- Step 5:** Conclude by either rejecting or not rejecting the null hypothesis at the given level of significance

# Exercise

---

**Exercise 6.** An insurance company is reviewing its current policy rates. When originally setting the rates they believed that the average claim amount was \$1,800. They are concerned that the true mean is actually higher than this, because they could potentially lose a lot of money. They randomly select 40 claims, and calculate a sample mean of \$1,950. Assuming that the standard deviation of claims is \$500, set  $\alpha = 0.05$  and test to see if the insurance company should be concerned.

[**Hint:** At 5% significance level, the critical value for a one tailed test from the table of z-scores = 1.645 ]

---

## Chapter 5: Scatter Plots and Correlations

---

# 5.1 Scatter Plot

## 5.1.1. Explanatory and Response Variables

A response variable measures an outcome of a study. An explanatory variable explains or influences changes in a response variable.

- A **response variable** is also called **dependent variable** and is generally denoted by **Y**
- An **explanatory variable** is also called **independent variable** and is generally denoted by **X**

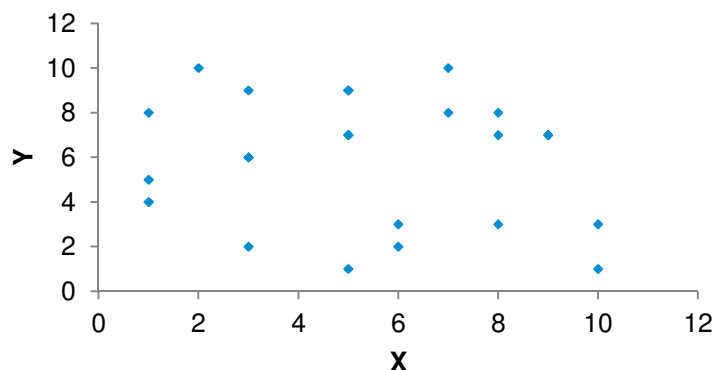
Analysis of relationship between two variables is referred to as **Bivariate Analysis**. To run this analysis, both variables are measured on the same individuals.

Relationships between two quantitative variables are best displayed graphically through a scatterplot

## 5.1.2. Plotting and Visualizing a Scatter Plot

A scatterplot shows the relationship between two quantitative variables measured on the same individuals.

- The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis
- Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual

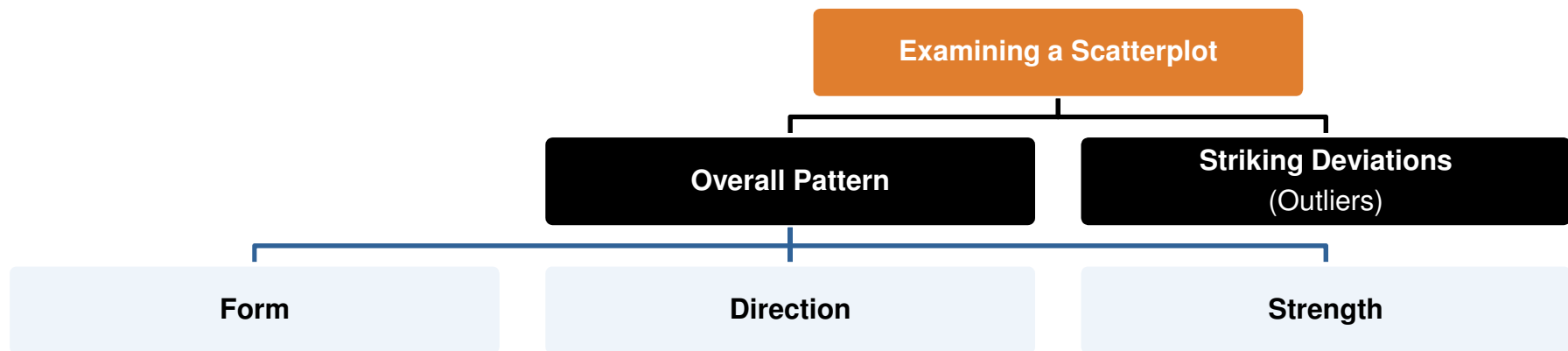


- ✓ Always plot the explanatory variable on the horizontal axis (the x axis) of a scatterplot
- ✓ If there is no explanatory-response distinction, either variable can go on the horizontal axis

---

### 5.1.3. Examining a Scatterplot

Look for the **overall pattern** and for **striking deviations** from that pattern



- **Form:** Linear relationships, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and clusters are other forms to watch for.
- **Direction:** If the relationship has a clear direction, there is either positive association (high values of the two variables tend to occur together) or negative association (high values of one variable tend to occur with low values of the other variable)
- **Strength:** The strength of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line

## 5.2 Correlation

### 5.2.1. Pearson's Correlation Coefficient

Pearson's Correlation Coefficient ( $r$ ) measures the strength and direction of the linear association between two quantitative variables  $X$  and  $Y$ . Although correlation can be calculated for any scatterplot, Pearson's correlation coefficient measures only straight-line relationships.

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \in [-1, +1]$$

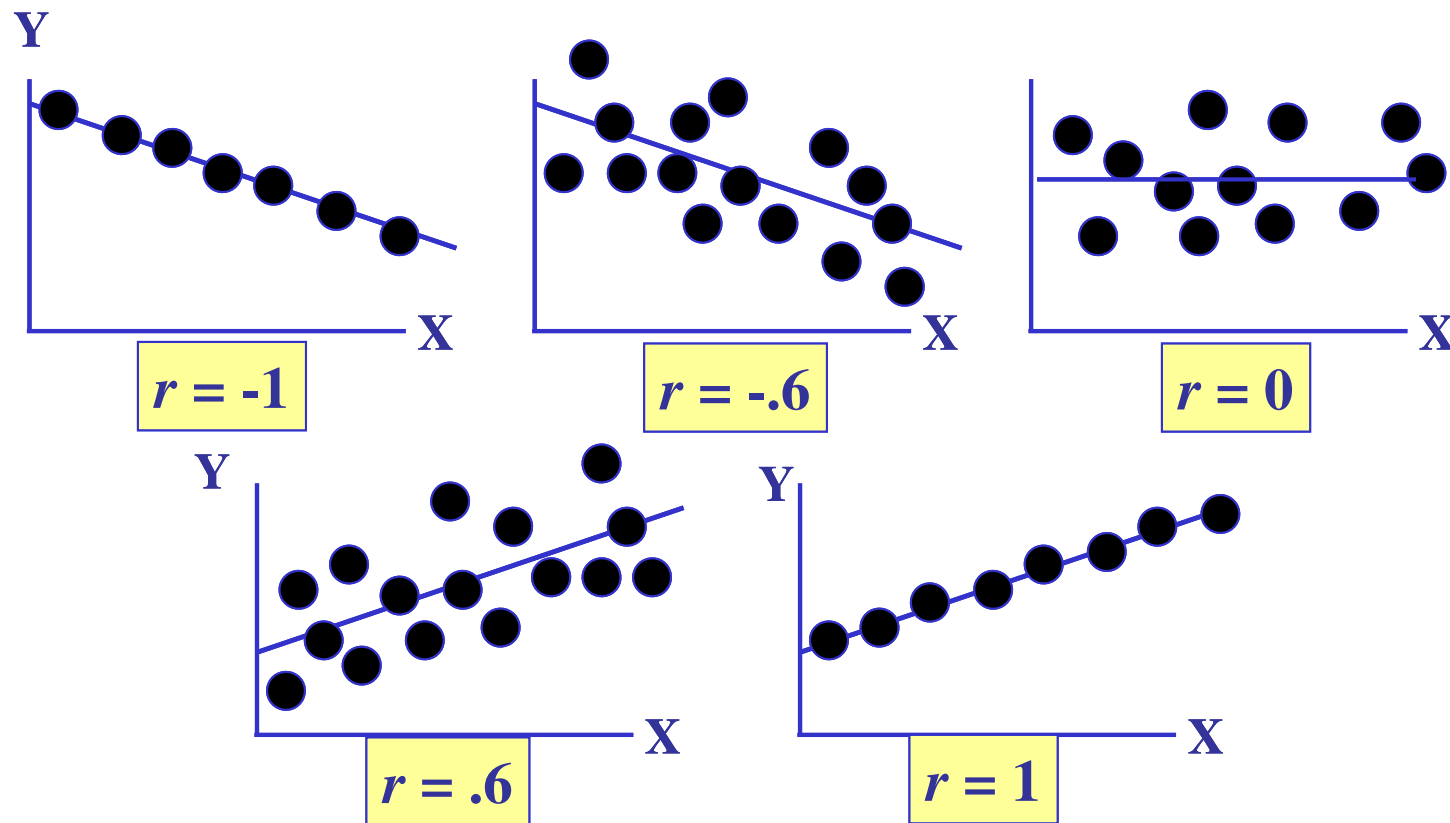


**SAS Tip**

**PROC CORR**  
(Use OUTP Option)

- Correlation indicates the **direction** of a linear relationship by its sign
  - $r > 0$  for a positive association
  - $r < 0$  for a negative association
- Correlation indicates the **strength** of a linear relationship by magnitude of its absolute value
  - Correlation ranges from -1 to +1
  - The Closer to -1, the Stronger the Negative Linear Relationship
  - The Closer to +1, the Stronger the Positive Linear Relationship
  - The Closer to 0, the Weaker Any Linear Relationship
  - $r = \pm 1$  means perfect correlation, which occurs only when the points on a scatterplot lie exactly on a straight line
- Correlation ignores the distinction between explanatory and response variables
- The value of  $r$  is not affected by changes in the unit of measurement of either variable
- Correlation is not resistant, so outliers can greatly change the value of  $r$

## Graphical Illustration





## Example

	A	B	C	D	E	F	G
1	X	Y	$X_i - X_{\text{mean}}$	$Y_i - Y_{\text{mean}}$	$(X_i - X_{\text{mean}}) * (Y_i - Y_{\text{mean}})$	$(X_i - X_{\text{mean}})^2$	$(Y_i - Y_{\text{mean}})^2$
2	101.0	99.2	24.7	(35.3)	(871.6)	609.5	1246.5
3	100.1	99.0	23.8	(35.5)	(844.6)	565.9	1260.7
4	100.0	100.0	23.7	(34.5)	(817.4)	561.1	1190.7
5	90.6	111.6	14.3	(22.9)	(327.3)	204.2	524.7
6	86.5	122.2	10.2	(12.3)	(125.4)	103.8	151.4
7	89.7	117.6	13.4	(16.9)	(226.3)	179.2	285.8
8	90.6	121.1	14.3	(13.4)	(191.5)	204.2	179.7
9	82.8	136.0	6.5	1.5	9.7	42.1	2.2
10	70.1	154.2	(6.2)	19.7	(122.3)	38.6	387.9
11	65.4	153.6	(10.9)	19.1	(208.4)	119.1	364.6
12	61.3	158.5	(15.0)	24.0	(360.2)	225.4	575.7
13	62.5	140.6	(13.8)	6.1	(84.2)	190.8	37.1
14	63.6	136.2	(12.7)	1.7	(21.5)	161.6	2.9
15	52.6	168.0	(23.7)	33.5	(794.2)	562.2	1121.9
16	59.7	154.3	(16.6)	19.8	(328.8)	276.0	391.8
17	59.5	149.0	(16.8)	14.5	(243.7)	282.6	210.1
18	61.3	165.5	(15.0)	31.0	(465.3)	225.4	960.6
19					(6023.1)	4551.5	8894.2

$$r = (-6023.1) / (4551.5 * 8894.2)^{1/2}$$

$$r = -95\%$$

### 5.2.2. Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient (rho), is a non-parametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function.

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \text{ where } d_i = x_i - y_i$$



**SAS Tip**

**PROC CORR**  
(Use OUTS Option)

#### Example:

	A	B	C	D	E
1	Candidate	Judge A's Ranking	Judge B's Ranking	d	d <sup>2</sup>
2	A	1	2	-1	1
3	B	2	1	1	1
4	C	3	4	-1	1
5	D	4	3	1	1
6	E	5	6	-1	1
7	F	6	7	-1	1
8	G	7	5	2	4
9	Sum				10

$$\rho = 1 - [6(10) / 7(49-1)]$$

$$\rho = 82.14\%$$

---

# Appendix

---

# A.1 List of Some Commonly Used Tests

---

## Statistical Tests

[Back to Main Slide](#)

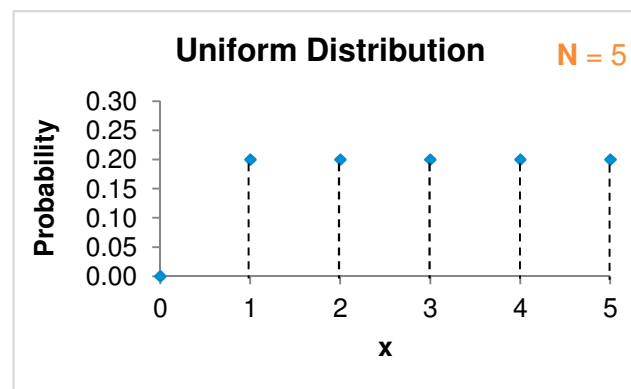
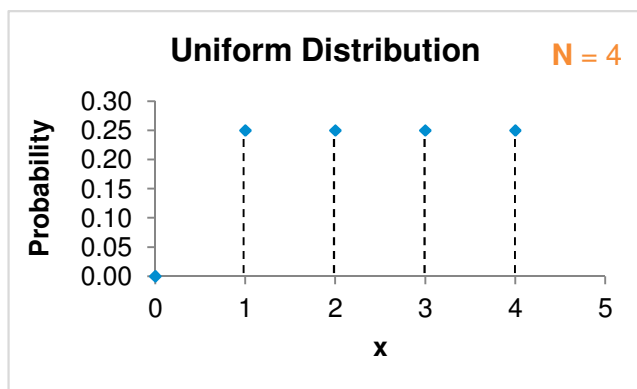
Investigation of significance of

- Difference between an assumed population mean  $\mu_0$  and a sample mean  $\bar{x}$  ,
  - When the population variance is known [ Z-test ]
  - When the population variance is unknown [ t-test ]
- Difference between two sample means, one from each population,
  - When the population variances are known and equal [ Z-test ]
  - When the population variances are known and unequal [ Z-test ]
  - When the population variances are unknown but equal [ t-test ]
  - When the population variances are unknown and unequal [ t-test ]
  - When observations for the two sample are obtained in pairs [ t-test ]
- Difference between an assumed population proportion and an observed sample proportion [ Z-test ]
- Difference between two sample proportions, one from each population [ Z-test ]
- Difference between two counts [ Z-test ]
- Difference between an assumed population variance and a sample variance [  $\chi^2$ -test ]
- Difference between two sample variances, one from each population [ F-test ]
- A variable in Regression Model (i.e. difference between a regression coefficient and zero) [ t-test ]
- Overall regression model [ F-test ]

# A.2 Examples of Discrete Distributions

## A.2.1. Discrete Uniform Distribution

[Back to Main Slide](#)



### Probability Mass Function

$$P(X = k) = \frac{1}{N}, \quad k = 1, 2, \dots, N$$

### Main Properties

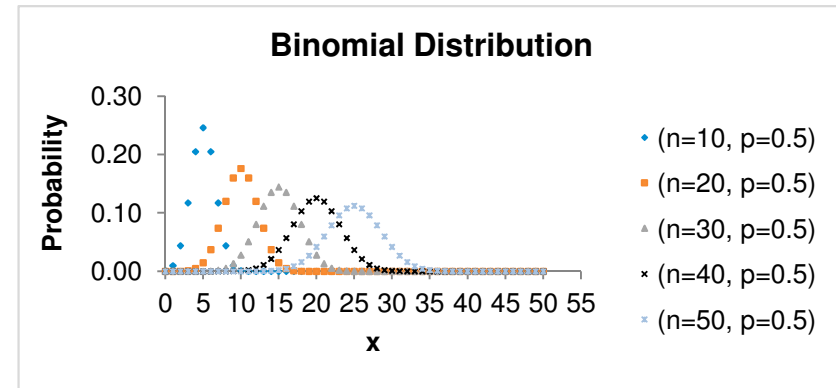
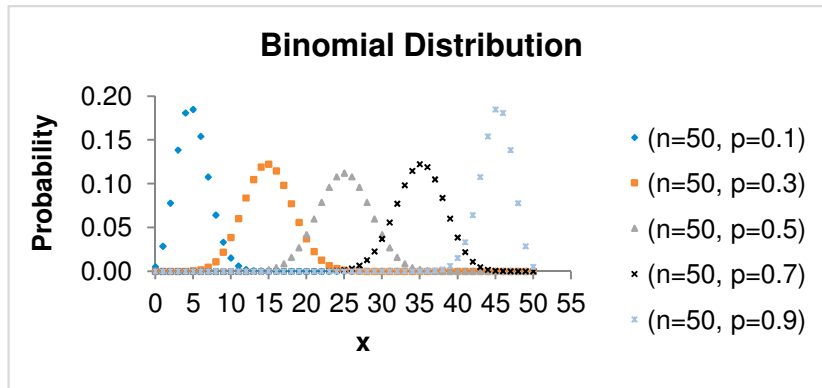
- This distribution is used to model experimental outcomes which are “equally likely”

### Applications

- Tossing of a fair and unbiased die. For the given sample space  $\{1, 2, 3, 4, 5, 6\}$ , each number occurs with a probability of  $1/6$

## A.2.2. Binomial Distribution

[Back to Main Slide](#)



### Probability Mass Function

$$P(X = k) = {}^n C_k p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n$$

**n** : Number of trials  
**p** : Probability of success on a single trial

### Main Properties

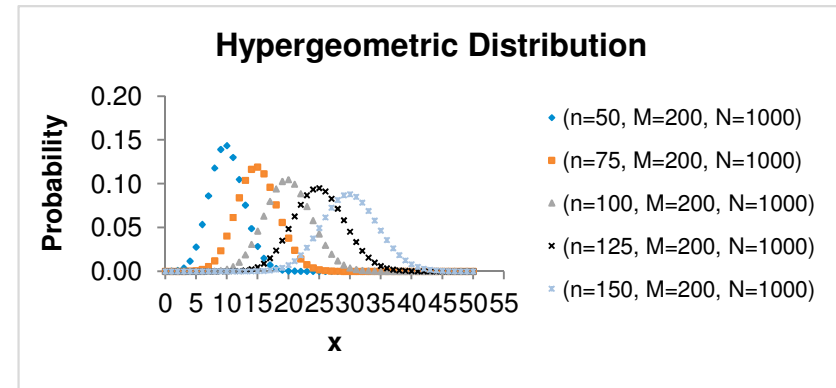
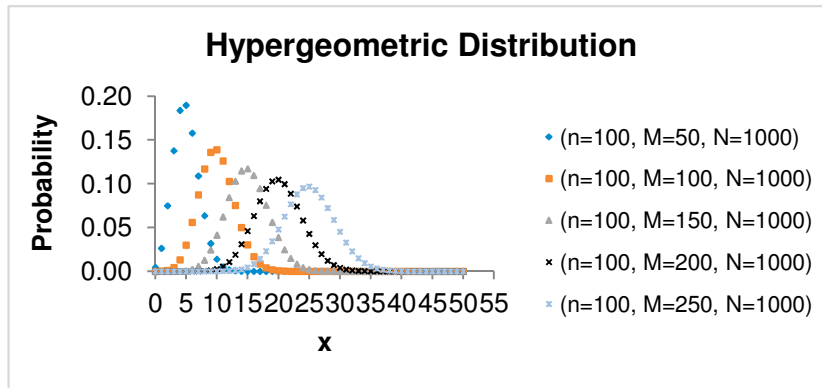
- Binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial, labeled as 'success' and 'failure'
- Mean =  $n p$  and variance =  $n p (1-p)$
- The binomial distribution is probably the most commonly used discrete distribution

### Applications

- Customer Retention (Identification of Attrition Cases)
- Fraud Detection (Identification of Defaulters)

## A.2.3. Hypergeometric Distribution

[Back to Main Slide](#)



### Probability Mass Function

$$P(X = k) = \frac{{}^M C_k {}^{N-M} C_{n-k}}{{}^N C_n}, \quad L \leq k \leq U, \text{ where } L = \max\{0, M-N+n\} \text{ and } U = \min\{n, M\}$$

### Main Properties

- The Hypergeometric (n, M, N) distribution models the number of items of a particular type that are there in a sample of size n where that sample is drawn from a population of size N of which M are also of that particular type
- Mean =  $n (M / N)$  and Variance =  $n (M / N) [1 - (M / N)] [(N-n) / (N-1)]$

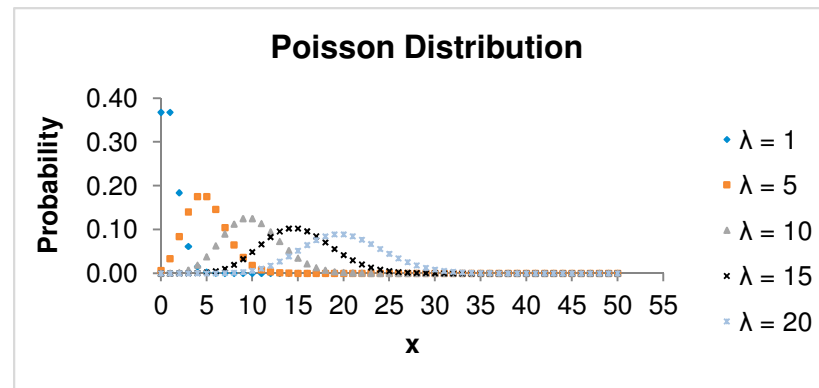
**N** : Number of items in population  
**M** : Number of defective items in population  
**n** : Number of items in a sample  
**k** : Number of defective items in the sample

### Applications

- Sampling without replacement
- From a lot of N items with M defective pieces, what is the probability of getting k defective items by the customer if he purchases n units?

## A.2.4. Poisson Distribution

[Back to Main Slide](#)



### Probability Mass Function

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

λ : The shape parameter which indicates the average number of events in the given time interval

### Main Properties

- Poisson distribution is used to model the number of events occurring within a given time interval
- Mean = Variance = λ
- Poisson distribution is very commonly used to model count data

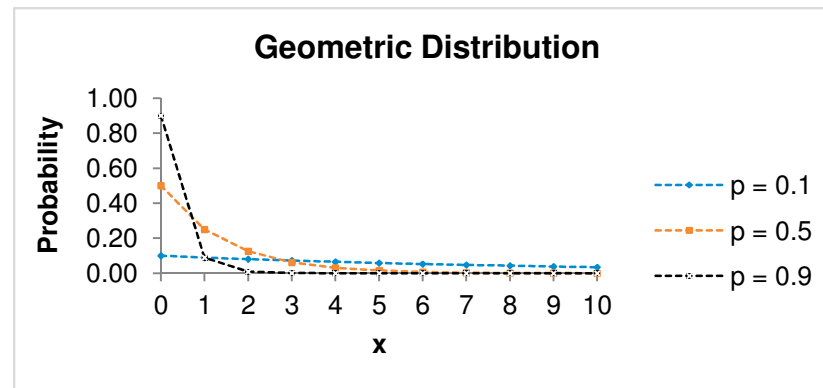
### Applications

- Count of phone calls arriving at a call centre per minute
- Count of insurance claims made by customers in a given period of time



## A.2.5. Geometric Distribution

[Back to Main Slide](#)



### Probability Mass Function

$$P(X = k) = (1 - p)^k p, \quad k = 0, 1, 2, \dots$$

**p** : Probability of success on each trial

### Main Properties

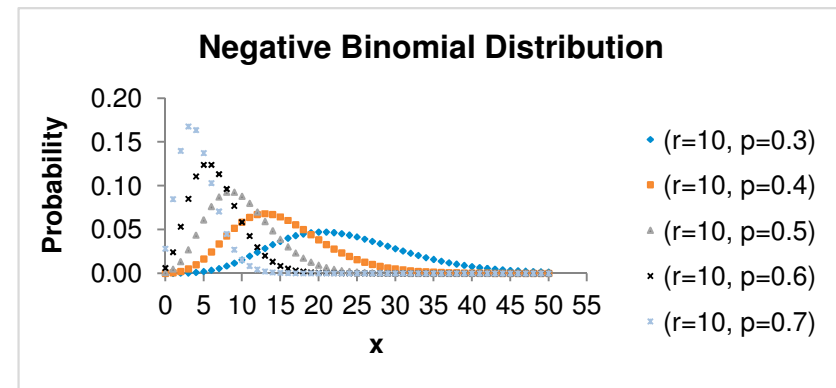
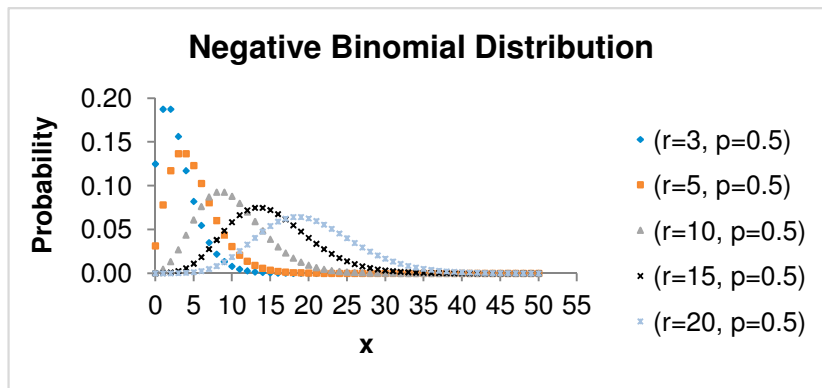
- Geometric distribution is used for modeling number of failures until the first success. It expresses the probability of exactly k failures until the first success to occur or equivalently, the probability that exactly (k + 1) trials are required to get the first success
- Mean =  $(1-p) / p$  and Variance =  $(1-p) / p^2$

### Applications

- Number of phone calls required before making a sale
- Number of dry wells an oil company will drill in a particular area before getting an oil-producing well
- Number of proposals to get a 'Yes'

## A.2.6. Negative Binomial Distribution

[Back to Main Slide](#)



### Probability Mass Function

$$P(X = k) = {}^{r+k-1}C_k p^r (1-p)^k, \quad k = 0, 1, 2, \dots; 0 < p < 1$$

**p** : Probability of success on each trial  
**r** : Threshold number of successes  
**k** : Number of failures until the  $r^{\text{th}}$  success

### Main Properties

- Negative Binomial distribution gives the probability of observing k failures before the  $r^{\text{th}}$  success or equivalently, probability that  $k+r$  trials are required until the  $r^{\text{th}}$  success to occur
- Mean =  $r(1-p)/p$  and Variance =  $r(1-p)/p^2$

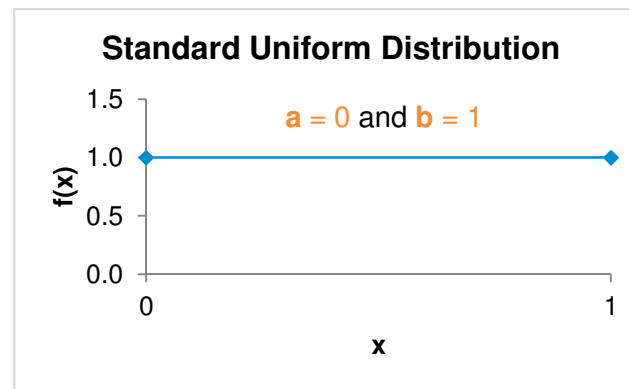
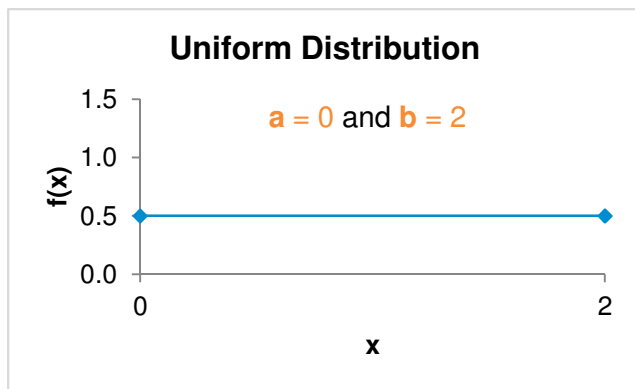
### Applications

- Games (what's the probability that a basket ball player makes his 3<sup>rd</sup> free throw on his 5<sup>th</sup> shot?)
- Marketing (what's the probability that a door-to-door salesman sells the last candy bar at the 10<sup>th</sup> house?)

# A.3 Examples of Continuous Distributions

## A.3.1. Continuous Uniform Distribution

[Back to Main Slide](#)



### Probability Density Function

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

### Main Properties

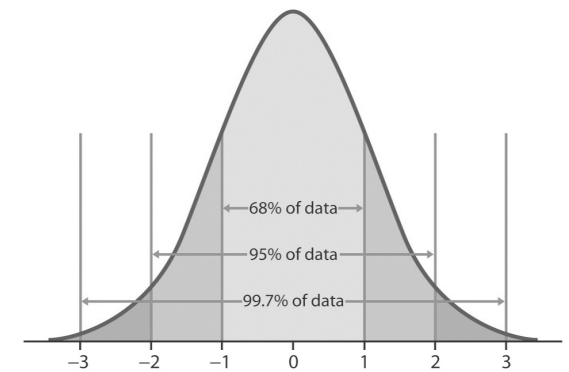
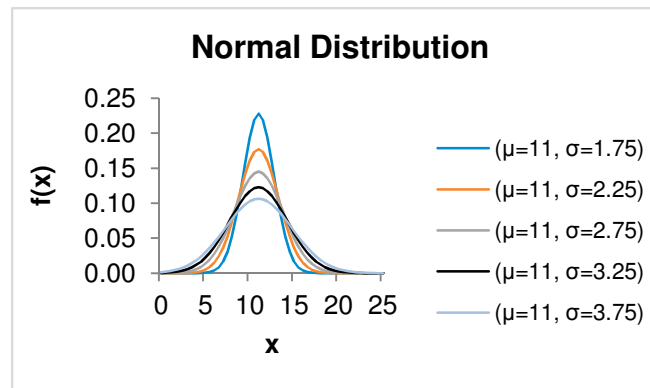
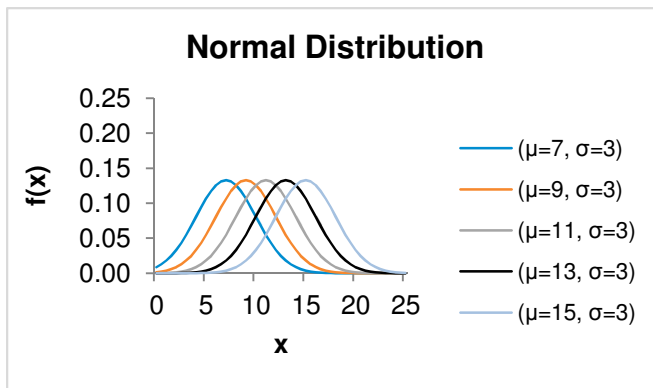
- The uniform distribution defines equal probability over a given range for a continuous distribution
- 'a' is the location parameter and 'b-a' is the scale parameter
- In case of a standard uniform distribution,  $a = 0$  and  $b = 1$

### Applications

- Generation of random numbers: Almost all random number generators generate random numbers on the (0,1) interval

## A.3.2. Normal Distribution

[Back to Main Slide](#)



### Probability Density Function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

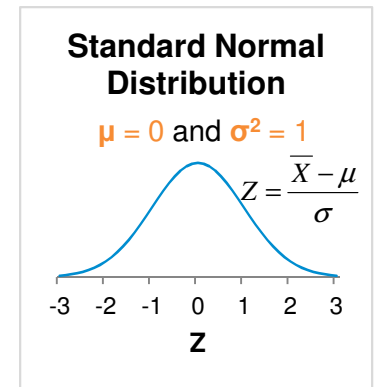
$\mu$  : Mean  
 $\sigma$  : Standard Deviation

### Main Properties

- It is symmetrical around its mean value and is also called the 'bell-shaped' curve
- 68%, 95% and 99.7% of area under the normal curve lies within  $\mu \pm \sigma$ ,  $\mu \pm 2\sigma$ , and  $\mu \pm 3\sigma$  respectively
- Normal distribution with mean = 0 and variance = 1 is called standard normal distribution
- Central Limit Theorem:** Let  $X_1, X_2, \dots, X_n$  denote  $n$  independent random variables, all of which have same PDF with mean =  $\mu$  and variance =  $\sigma^2$ . Let  $\bar{X} = \sum X_i / n$ . Then as  $n$  increases indefinitely (i.e. as  $n \rightarrow \infty$ ),  $\bar{X}$  approaches the normal distribution with mean =  $\mu$  and variance =  $\sigma^2 / n$

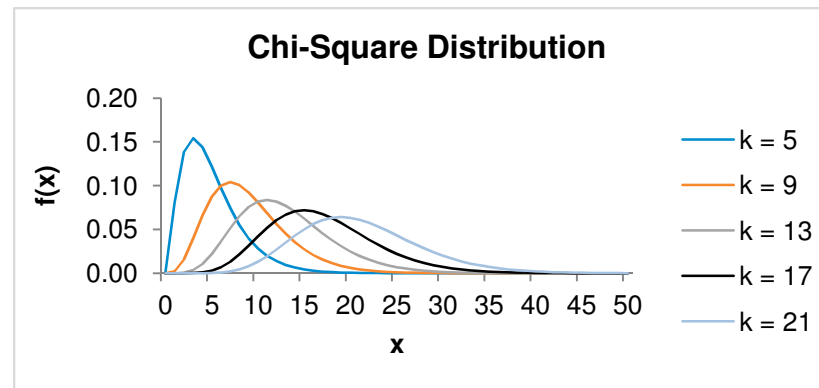
### Applications

- Any process yielding values that tend to be symmetric around a central value (the mean) generally follow a Normal Distribution like height of individuals and marks of candidates in an entrance exam



### A.3.3. Chi-Square ( $\chi^2$ ) Distribution

[Back to Main Slide](#)



#### Probability Density Function

$$f(x) = \frac{1}{2^{k/2} \Gamma(k/2)} e^{-\frac{x}{2}} x^{\frac{k}{2}-1}, \quad x > 0, k > 0, \Gamma \text{ is the Gamma Function : } \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \quad \mathbf{k} : \text{Degrees of freedom}$$

#### Main Properties

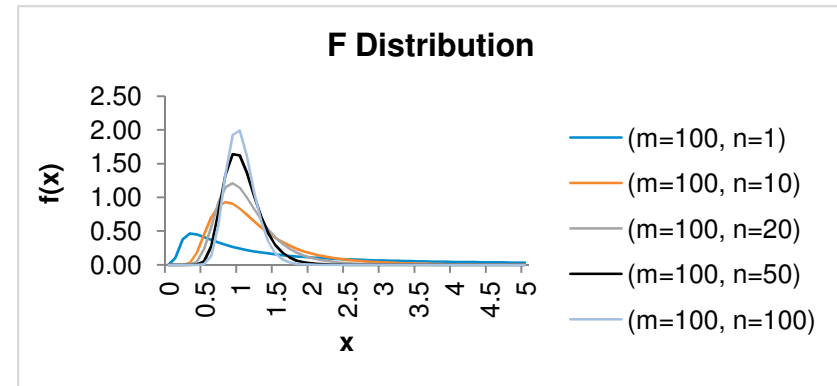
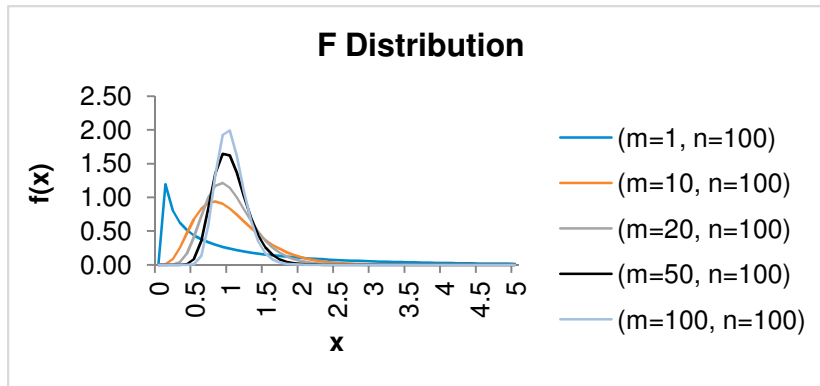
- If  $Z_1, Z_2, \dots, Z_k$  are independent standardized normal variables, then  $Z = \sum Z_i^2$  is said to possess the  $\chi^2$  distribution with  $df = k$
- The  $\chi^2$  distribution is a skewed distribution, the degree of skewness depending on the  $df$ . For comparatively few  $df$ , the distribution is highly skewed to the right; but as the  $df$  increases, the distribution becomes increasingly symmetrical
- Mean =  $k$  and Variance =  $2k$

#### Applications

- Hypothesis Testing (e.g. to test if a sample variance is equal to an assumed population variance)

### A.3.4. F Distribution

[Back to Main Slide](#)



#### Probability Density Function

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \times \frac{\left(\frac{m}{n}\right)^{m/2} x^{\frac{m}{2}-1}}{\left(1 + \frac{mx}{n}\right)^{\frac{m+n}{2}}}, \quad m > 0, n > 0, x > 0, \Gamma \text{ is the Gamma Function : } \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

**m** : DF of  $\chi^2$  random variable in the numerator  
**n** : DF of  $\chi^2$  random variable in the denominator

#### Main Properties

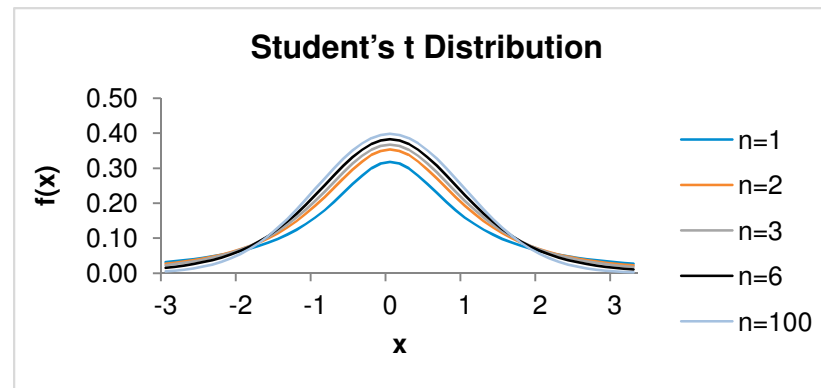
- If  $X_1$  and  $X_2$  are independently distributed chi-square variables with  $m$  and  $n$  df respectively, then  $F = [(X_1 / m) / (X_2 / n)]$  follows (Fisher's) F distribution with  $m$  and  $n$  df
- Like  $\chi^2$  distribution, F distribution is skewed to the right. But as  $m$  &  $n$  become large, F distribution approaches normal distribution

#### Applications

- Hypothesis Testing (e.g. Test of equality of variance of two populations, ANOVA test etc)

### A.3.5. Student's t Distribution

[Back to Main Slide](#)



#### Probability Density Function

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \times \left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \quad -\infty < x < \infty, \quad n \geq 1, \quad \Gamma \text{ is the Gamma Function : } \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt \quad \text{ } n : \text{DF of } \chi^2 \text{ random variable in the denominator}$$

#### Main Properties

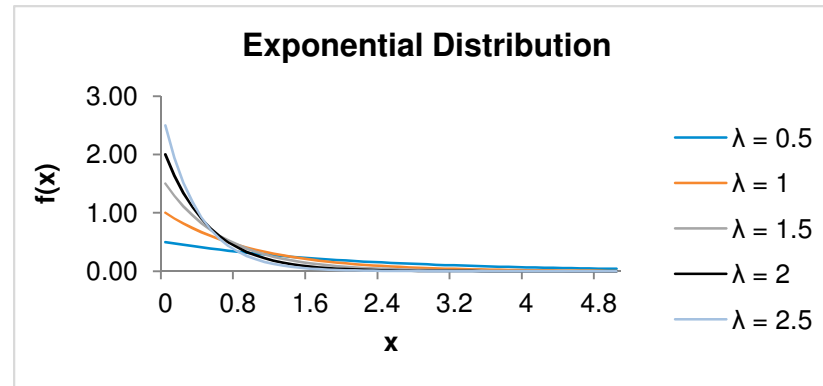
- Let Z and S be independent random variables such that  $Z \sim N(0,1)$  and  $n S^2 \sim \chi^2_n$ . The distribution of  $t = Z / S$  is called Student's t distribution with df = n. As df increases, t distribution approximates the normal distribution.
- Mean = 0 for  $n > 1$ , Variance =  $n / (n-2)$  for  $n > 2$ , Median = 0, Skewness = 0

#### Applications

- Hypothesis Testing (e.g. Estimation of population parameters when sample size is small or when population variance is unknown)

## A.3.6. Exponential Distribution

[Back to Main Slide](#)



### Probability Density Function

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0$$

### Main Properties

- The exponential distribution is used to model a random variable with mean  $1 / \lambda$ , that represents the waiting time until the first even to occur, where events are generated by a Poisson process with mean  $\lambda$  (i.e. events occur continuously and independently at a constant average rate  $\lambda$ )
- Mean =  $1 / \lambda$  and Variance =  $1 / \lambda^2$
- In case of a standard exponential distribution,  $\lambda = 1$

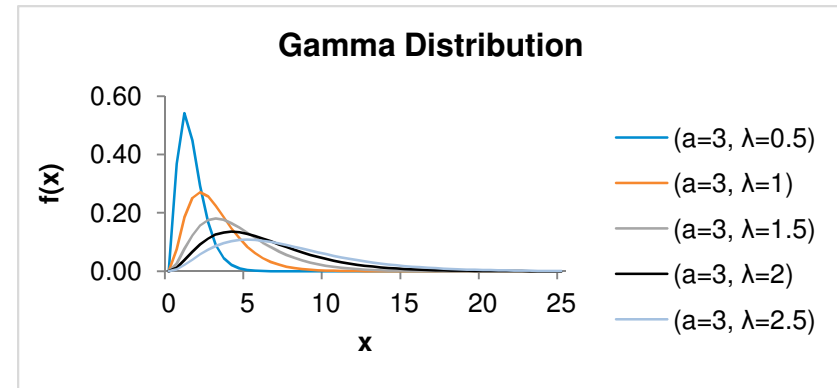
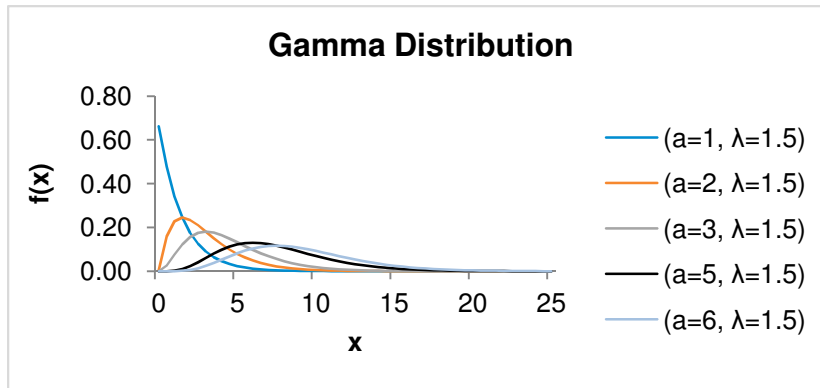
### Applications

- The exponential distribution is primarily used in reliability applications



## A.3.7. Gamma Distribution

[Back to Main Slide](#)



### Probability Density Function

$$f(x) = \frac{\lambda^a e^{-\lambda x} x^{a-1}}{\Gamma(a)}, \quad x > 0, \lambda > 0, a > 0$$

### Main Properties

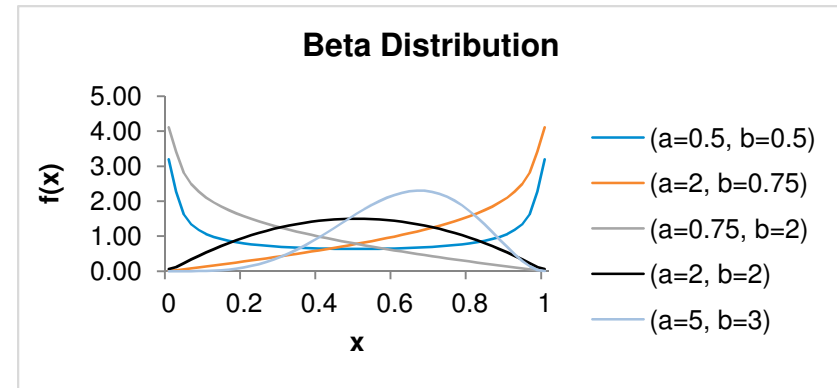
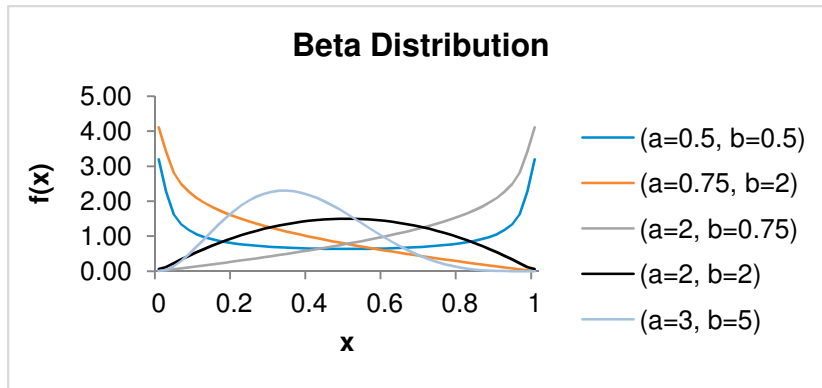
- The gamma distribution can be viewed as a generalization of the exponential distribution. The gamma random variable represents the waiting time until the  $a^{\text{th}}$  event to occur.
- Mean =  $a / \lambda$  and Variance =  $a / \lambda^2$

### Applications

- For modeling size of insurance claims
- For modeling size of rainfalls

## A.3.8. Beta Distribution

[Back to Main Slide](#)



### Probability Density Function

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \quad 0 < x < 1, a > 0, b > 0$$

### Main Properties

- a and b are shape parameters. The density curve is
  - U-shaped when  $a < 1$  and  $b < 1$
  - Symmetric about 0.5 when  $a = b > 1$
  - J-shaped when  $(a-1)(b-1) < 0$
  - Unimodal for all other values of a and b
- Mean =  $a / (a + b)$  and Variance =  $ab / [(a + b)^2(a + b + 1)]$

### Applications

- For modeling events which are constrained to take place within an interval defined by a minimum and maximum value